# UNIVERSITY OF GÖTEBORG

## Department of Statistics

RESEARCH REPORT 1983:2

ISSN - 00349-8034

# ANALYS AV KATEGORISKA DATA

En metodstudie i anslutning till statsvetenskaplig forskning.

av

SVEN ERIKSSON

**Statistiska institutionen**
**Göteborgs Universitet**
**Viktoriagatan 13**
**S 411 25 Göteborg**
**Sweden**

# ANALYS AV KATEGORISKA DATA

## En metodstudie i anslutning till statsvetenskaplig forskning*

Sven Eriksson

* Denna forskning har genomförts med stöd av Humanistisk-samhällsvetenskapliga forskningsrådet (Anslagen F381/80 och F425/81)

# INNEHÅLLSFÖRTECKNING

# BAKGRUND

## Kommunforskningsprogrammet

Inriktningen av denna rapport har präglats av författarens erfarenheter från arbetet med urvalsplaner för delundersökningarna avseende medborgare, politiker och tjänstemän inom forskningsprogrammet 1979-1981 för utvärdering av kommunsammanslagningsreformen, av allmänna diskussioner om analysproblem under planeringsstadiet av forskningsprogrammet samt av studium av de olika projektens slutrapporter. I forskningsprogrammet deltog forskare från samtliga statsvetenskapliga institutioner i Sverige med undantag för Uppsala samt forskare från kulturgeografiska institutionen i Lund.
Författaren har redovisat urvalsplaner och estimation av populationskarakteristikor samt diskuterat vissa allmänna metodfrågor i tidigare rapporter ([ 4 ], [26 ] bilagor samt [28 ] kap. 12 ).

## Den statistiska analysens betingelser

Kommunforskningsprogrammets delundersökningar, liksom flertalet samhällsvetenskapliga undersökningar som genomföres i surveyform, präglas ur statistisk synpunkt främst av nedanstående förhållanden.

a. Det finns dubbla syften, ett beskrivande och ett förklarande. Man önskar beskriva populationen och olika delgrupper med hjälp av enkla mått: relativa och absoluta frekvenser av olika egenskaper och kombinationer av egenskaper samt medelvärden och totalvärden
Man förklarar egenskaper, attityder, politiskt handlande och andra företeelser med hjälp av individ- och miljövariabler (i detta fall kommunvariabler). Ofta ingår även attitydvariabler bland förklaringsvariablerna.
Häri finns en konflikt mellan olika önskemål inbyggd.
Det deskriptiva syftet uppnås oftast bäst med ett komplext urvalsförfarande, i många fall med kraftigt varierande inklusionssannolikheter för olika element i populationen.
Ur sambandsanalysens synpunkt är det visserligen av intresse att få en representation i urvalet garanterad för alla viktigare delgrupper i populationen. Detta kan uppnås med hjälp av ett komplext urvalsförfarande.

Mot ett sådant urvalsförfarande talar emellertid mycket starkt det faktum att det till allra största delen saknas statistisk teori för sambandsanalys baserad på komplexa urval.
Kommunforskningsprogrammet bygger på komplexa urvalsförfaranden. Skälen härför redovisas i den ovan givna referensen [ 4 ].

b. De verkliga sambanden är av komplicerad art och det finns i allmänhet ett stort antal presumptiva förklaringsvariabler till en viss responsvariabel. Även om det finns kunskaper om sakförhållandena saknas i de flesta fall etablerad saklogisk teori som stöd vid konstruktion av statistiska modeller.

c. De variabler som observeras är i många fall av kvalitativ typ. I kommunforskningsprogrammet dominerar denna variabeltyp totalt i flertalet delundersökningar.

d. Urvalsplanerna är ofta komplexa med urvalssannolikheter som varierar kraftigt mellan olika element i populationen.

Genomförandet av den statistiska analysen

Den deskriptiva delen av kommunforskningsprogrammet ger upphov till vissa komplicerade estimationsproblem. Det är emellertid analysen av samband mellan kvalitativa variabler som varit stötestenen för forskare inom de olika projekten.
I vissa projekt bygger analysen nästan uteslutande på olika tabelleringar av materialet. I andra fall kompletteras denna med beräkningar av enkla korrelationer och regressionslinjer. Vissa författare, främst Westerståhl och Johansson [29] går längre och utnyttjar variansanalys med flera faktorer, och multipel linjär regressionsanalys samt i enstaka fall diskriminantanalys och faktoranalys. Ingen av de använda metoderna är emellertid utvecklade för analys av kvalitativa responsvariabler.
Skälet till att man valt ovan angivna metoder är naturligtvis att den multivariata teorin och metodiken för att studera samvariation och samband mellan kvantitativa variabler sedan länge är välutvecklad. Den är också allmänt känd och utnyttjad bland empiriskt inriktade forskare.

Tidigare har det däremot endast funnits en knapphändig och föga kraftfull metodik för multivariat analys av kvalitativa variabler. Under senare år har det emellertid försiggått en stark utveckling inom det senare området. Den torde dock vara ganska okänd utanför fackstatistikernas krets.

## Modellbegreppet

Genomgående har de statistiska metoderna vid analysen av kommunundersökningarna använts utan att man specificerar en statistisk modell. Statistisk analys utan modellspecifikation leder lätt till att en standardmodell regelmässigt utnyttjas och att resultaten tolkas utan att modellens validitet undersökts. Det föreligger också en uppenbar risk för val av mindre lämpliga eller felaktiga metoder.

## Disposition

Mot denna bakgrund är det angeläget att inledningsvis kortfattat beröra

- modellbegreppets roll i statistisk analys
- statistiska inferensens begränsningar i explorativa undersökningar
- olika undersökningstyper (experiment, urval, tidsserieobservation) som i allmänhet leder till olika modellspecifikationer
- ett komplext urvalsförfarandes inverkan på analys av statistiska modeller

Därefter behandlas

- specifikation och skattning av linjära regressionsmodeller
- användning av sådana modeller för analys av binära responsvariabler inom kommunforskningsprogrammet
- icke-linjära regressionsmodeller för binära responsvariabler

Senare delen av framställningen ägnas sedan åt

- presentation av nyare modeller och metoder för analys av samvariation mellan kategoriska (kvalitativa och klassindelade kvantitativa) variabler i flerdimensionella kontingenstabeller
- analys av olika frågeställningar inom kommunforskningsprogrammet på basis av sådana modeller.

# 1. STATISTISKT ANALYSARBETE

## 1.1 Modellbegreppets betydelse

En statistisk modell utgör en förenklad och matematisk formaliserad bild av ett visst fenomen. Modellen beskriver dels regelbundenheten eller den systematiska strukturen hos detta fenomen dels de slumpmässiga variationerna. Modellkonstruktion är en fråga om balansgång. Det måste göras en avvägning mellan modellens enkelhet och dess förmåga av vid ökad matematisk komplexitet på ett adekvat sätt återge mer eller mindre markerade särdrag som det aktuella fenomenet i verkligheten uppvisar.
Modellen utgör en länk mellan verklighet och data. Utan denna länk skulle det vara mycket svårt att vinna något större mått av kunskap om det fenomen som analyseras. På basis av datamaterialet uppskattas olika storheter (parametrar) i den modell som formularats. I nästa steg skall man på basis av slutsatserna om modellen dra slutsatser om det fenomen i verkligheten som studeras.
Slutsatserna i en studie står och faller med modellens giltighet. Det är därför nödvändigt att i allt statistiskt analysarbete ställa modellen i fokus. Modellen måste beskrivas noggrant (modellspecifikation). Det är nödvändigt att pröva modellens giltighet och att redovisa resultatet av denna prövning.

## 1.2 Konfirmativa undersökningar visavi dataanalys

Vid diskussion av användning och tolkning av statistiska modeller är det av intresse att skilja mellan två huvudsituationer. I det ena fallet har man på förhand goda kunskaper om vilka faktorer som påverkar en viss responsvariabel och kan stödja sig på etablerad saklogisk teori inom området. I det andra fallet saknas sådan teori och man kan på förhand kanske bara ange ett stort antal presumptiva förklaringsvariabler till en viss responsvariabel. Ofta finns förklaringsvariabler som man inte tänkt på eller som man inte kan erhålla observationer på.

Användningen av statistiska modeller och framför allt möjligheterna att dra slutsatsen av undersökningsresultaten varierar starkt mellan dessa båda fall som kan karakteriseras som konfirmativa undersökningar och explorativa undersökningar (dataanalys). Arbetsgången i de båda undersökningstyperna kan sammanfattas i figur 1.

Figur 1.1 Utformning av statistiska modeller i konfirmativa undersökningar och vid dataanalys (explorativa undersökningar).

| Konfirmativ undersökning | Dataanalys |
|---|---|
| Sakproblem formuleras och översätts till statistiska problem. | |
| Statistisk undersökningsplan utformas. | |
| Statistisk modell specificeras på basis av förhandskunskaper och etablerad teori inom sakområdet. | |
| Data insamlas enligt undersökningsplanen. | |
| Modellen skattas. Modellrelevans kontrolleras med hjälp av residualmönster och test. | På basis av förhandskunskaper och mönster i datamaterialet specificeras, prövas och modifieras en svit av modeller. |
| Parameterskattningar och osäkerhetsmått presenteras. | Parameterskattningar och osäkerhetsmått presenteras för en eller flera modeller som ger plausibla förklaringar mot bakgrund av förhandskunskaperna och som ger god anpassning till data. |
| Slutsatser dras om sakproblemet. | Vissa slutsatser kan dras om sakproblemet. |

## 1.3 Dataanalys och statistiska test

I den konfirmativa undersökningstypen är den statistiska modellen specificerad på förhand. Om den saklogiska teorin verkligen är väletablerad kommer i normalfallet den statistiska modellen att betraktas som adekvat efter att den konfronterats med det observerade datamönstret. Den statistiska inferensens formella mått på osäkerhet vid test och konfidensintervall för modellparametrar kommer då att vara tillämpbara.

Vid dataanalys (explorativa undersökningar) specificeras och modifieras en rad modeller med ledning av det observerade datamönstret. Som hjälpmedel vid denna modellbyggnadsprocess används vanligen förutom grafer över residualmönster olika statistiska test. Eftersom modellens utformning styrs av det observerade datamönstret svarar emellertid den formella testnivån inte mot samma reella nivå. Ju fler modellvarianter som prövas och ju fler presumptiva förklaringsvariabler man har att välja sina förklaringsvariabler bland, desto större är naturligtvis möjligheten att finna en modell som står i god överensstämmelse med de speciella särdrag som av en ren tillfällighet uppträder i det observerade urvalet.

Dataanalys leder därför i de flesta fall till överanpassade modeller. Man får en skönmålning av hur väl modellen, som stegvis har mejslats ut och filats av för att beskriva egenheterna i detta speciella urval, beskriver verkligheten.

Det torde i samhällsvetenskapliga surveys vara ovanligt att man kan specificera sina modeller på förhand. Det är därför inte heller möjligt att på basis av resultatet från en enstaka undersökning etablera kausala modeller (av vissa författare kallade funktionella modeller). Det bästa man kan hoppas på är en god prediktionsmodell. Man bör om möjligt ändå ställa sådana krav i modellbyggnadsarbetet att ingenting motsäger att den slutliga modellen skulle kunna vara en god förklaringsmodell. Man får därför inte låta den serie av test som i allmänhet förekommer under konstruktionsarbetet för att nå fram till en slutmodell bli det helt avgörande för dennas utseende.

## 1.4 Undersökningsformer och modeller

Det är i denna framställning av intresse att särskilja de två undersökningsformerna experiment och tvärsnittsundersökningar av surveytyp. Experimentet utgör, den i de flesta samhällsvetenskapliga sammanhang uteslutna, idealformen som ger helt andra möjligheter att dra pålitliga slutsatser om orsak/verkanrelationer än de båda andra undersökningsformerna.

a. Vid ett experiment förfogar experimentatorn över ett antal försöksenheter. Denne fördelar själv försöksenheterna på de olika behandlingarna enligt en viss experimentplan. Med andra ord bestäms vilka kombinationer av värden på förklaringsvariablerna man skall observera en viss responsvariabel. Värdena på förklaringsvariablerna kan därför betraktas som på förhand givna konstanter. Vanligen tillses att förklaringsvariablerna göres ortogonala (oberoende), vilket innebär att för varje "nivå" på en av variablerna varieras de övriga enligt samma mönster.
Randomiseringen, dvs. det slumpmässiga valet av försöksenhet för varje enskild observation samt den slumpmässiga ordningsföljden mellan observationer för olika kombinationer av nivåer på de kontrollerade variablerna garanterar att de i experimentet ingående kontrollerade variablerna inte på ett systematiskt sätt samvarierar med eventuella andra variabler som påverkar responsvariabeln. Dessas inflytande, om det existerar något sådant, kommer därför endast att ta sig uttryck i form av större slumpvariation hos responsvariabeln. De stör därför inte de kontrollerade variablernas systematiska inverkan på responsvariabeln. Randomiseringen medför också, såvitt det inte förekommer speciella förhållanden, att de olika observationerna kan betraktas som oberoende av varandra.

De ur analys- och modellsynpunkt intressantaste egenskaperna hos ett experiment är att

- man vet tack vare randomiseringen att det endast är de experimentellt kontrollerade variablerna som kan ha haft en systematisk effekt (det är ju inte säkert att alla verkligen har en effekt) på responsvariabeln. Modellbyggnaden blir därför, relativt sett, ett ganska enkelt problem.
- förklaringsvariablerna har i experimentplanen vanligen gjorts ortogonala. De olika förklaringsvariablernas effekter blir därför inte hopblandade med varandra. Man kan också utan komplikationer studera interaktioner dvs. hur effekten av en förklaringsvariabel beror av de övriga förklaringsvariablernas nivåer.
- observationerna kan, tack vare randomiseringen, betraktas som statistiskt oberoende. Detta bidrar till att göra analysen av den statistiska modellen enkel.
- sannolikhetsmodellerna blir enkla. Observationernas variation för en viss kombination av nivåer på förklaringsvariablerna beskrivs av endimensionella fördelningar eftersom endast responsvariabeln är slumpmässig. (För en konkretisering se kap. 2.2.) Detta medför att det blir enkelt att härleda estimatorer enligt olika metoder. Det medför också att det inte är meningsfullt att beräkna mått på samvariation (såsom korrelationskoefficienter) mellan responsvariabel och olika förklaringsvariabler, ett tyvärr ibland förekommande komplement till övrig beskrivning av experimentutfall.

b. Samhällsvetenskaperna arbetar ofta med slumpmässiga urval av individer från en ändlig population. I vissa fall genomföres totalundersökningar men detta ändrar inte problemen vid modellbyggnaden

på annat sätt än att man slipper ta hänsyn till någon urvalsplan. En modell som specificeras för att beskriva det aktuella sakproblemet inkluderar m variabler. De m mätvärden som erhålles för en slumpmässigt vald individ betraktas som en observation på en m-dimensionell stokastisk variabel. Modellen som beskriver företeelsen kommer därför att utgöras av en m-dimensionell sannolikhetsfördelning.
Vid analys av surveydata har alla ovan uppräknade fördelar med experiment bytts mot extra komplikationer.

- Modellbyggnadsprocessen blir vansklig därför att det vanligen inte är möjligt att ange hur många och vilka variabler som bör ingå i den statistiska modellen.
- Variablerna i modellen samvarierar och dessutom föreligger en samvariation med variabler som ej är inkluderade i modellen. Om någon ej inkluderad variabel påverkar den aktuella företeelsen uppstår ofrånkomligen systematiska fel vid skattning av ingående modellvariablers effekter. Man får med andra ord en hopblandning av effekter ("confounding" enligt engelskspråklig terminologi).
- Det är inte möjligt att experimentellt manipulera förklaringsvariabler (oberoende av varandra) och avläsa effekten på responsvariabeln. Man kan därför principiellt sett aldrig avgöra om en konstaterad samvariation är ett uttryck för ett kausalt samband eller ej.
- Med många vanligt förekommande urvalsplaner (grupp- och flerstegsurval) blir observationerna statistiskt beroende.
- Ofta ges olika individer olika urvalssannolikheter i urvalsplanen.
- Sannolikhetsmodellerna blir komplicerade eftersom de utgörs av flerdimensionella fördelningar för beroende variabler. (Se kap. 2.2.)

- Ofta föreligger en ömsesidig påverkan mellan två eller flera variabler vilket ytterligare ökar modellens komplexitetsgrad.

Undersökningar baserade på tidsserier av observationer dras i allt väsentligt med samma principiella problem som tvärsnittsundersökningar i urvalsform. Man måste därutöver ta i beaktande att i tiden på varandra nära följande observationer mycket ofta måste betraktas som beroende.

## 1.5 Surveyanalytikerns dilemma

Surveys har som tidigare nämnts i allmänhet två syften: Beskrivning av populationen i termer av enkla mått och frekvenser samt analys av samband.

A. Medlet för att åstadkomma en effektiv deskription är ett effektivt utnyttjande av den förhandsinformation (de hjälpvariabler) som föreligger. Den statistiska teorin för deskriptiva undersökningar behandlar olika komplikationer såsom
   - ett komplext urvalsförfarande med utnyttjande av hjälpinformation
   - speciell estimation med utnyttjande av hjälpinformation
   - svarsbortfall
   - mätfel vid observationerna.

B. Medlet för att åstadkomma en effektiv analys av samband mellan variabler är framför allt ett utnyttjande av sannolikhetsmodeller.
   Den statistiska inferensteorin (allmän inferensteori och multivariat teori) förutsätter emellertid i stort sett utan undantag en idealiserad situation
   - obundet slumpmässigt urval (oberoende observationer)
   - inget bortfall av observationer
   - inga mätfel förekommer.

   Till stora delar är det fråga om en normalfördelningsteori eftersom observationerna antas komma från en- eller flerdimensionella normalfördelningar.

## 1.6 Målet för den statistiska inferensen

Inom den experimentella statistiken finns en väletablerad och principiellt sett enkel begreppsapparat. De vid experimentserien erhållna mätvärdena betraktas som genererade enligt en viss sannolikhetsmodell. Experimentet är i princip upprepbart under oförändrade betingelser och önskat antal observationer kan erhållas. Någon konkret population av enheter som är bärare av mätvärdena existerar ej utan det enskilda experimentet genererar mätvärdet. Försöksenheterna i sig är inte intressanta i annan mening än att de utgör ett medel att studera den bakomliggande mekanismen eller orsakssambandet vilket representeras av sannolikhetsmodellen. Den statistiska inferensen avser sannolikhetsmodellen och dess parametrar.

I samhällsvetenskapliga undersökningar studeras ofta ändliga populationer av enheter, exempelvis personer. Man önskar dra slutsatser om vilka faktorer som påverkar, och på vilket sätt dessa faktorer påverkar en viss företeelse (exempelvis en viss form av politisk aktivitet). Frågan är nu om dessa slutsatser primärt avser enheterna i den ändliga populationen med just den sammansättning den hade vid undersökningstillfället eller om slutsatserna avser det bakomliggande "orsakssystemet". Den ändliga populationen utgör i det senare fallet endast ett medel att studera orsakssystemet. Vilket av de båda synsätten som väljs har avgörande betydelse för tillvägagångssättet vid inferensen och kräver principiellt helt olika begreppsapparater. Om den ändliga populationen i sig är inferensens mål kan det för att ta ett enkelt exempel vara aktuellt att skatta en regressionskoefficient B definierad som

$$B = \frac{\sum_{1}^{N}(x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{1}^{N}(x_i - \bar{x})^2}$$

Bortsett från ett allmänt vagt krav på att samvariationen mellan x och Y bör beskrivas väl med hjälp av en rät linje föreligger inga speciella antaganden för "regressionsana-

lysen". Gängse krav på att estimatorerna bör vara väntevärdesriktiga (eller i varje fall konsistenta) leder till att man vid estimationen måste väga de olika observationerna med sina urvalssannolikheter.
Om däremot det bakomliggande orsakssystemet utgör målet för inferensen beskrivs detta system med hjälp av en superpopulationsmodell vilken utgör en modell av samma typ som den experimentella statistikens sannolikhetsmodeller. Man tänker sig nu att den ändliga populationens mätvärden genererats enligt denna modell. I stället för regressionskoefficienten B ovan skulle man nu vara intresserad av regressionskoefficienten $\beta$ i modellen

$$Y_i = \alpha + \beta x_i + \varepsilon_i$$

där slumptermerna $\varepsilon_i$ i enklaste fallet ges den experimentella standardmodellens egenskaper (väntevärde 0, konstant varians, oberoende av varandra och normalfördelning). Den stora frågan är nu om urvalets observationer vid estimationen skall vägas med urvalsenheternas urvalssannolikheter eller ej. Resultatet enligt det ena förfaringssättet kan skilja sig starkt från resultatet enligt det andra förfaringssättet. Problemet ställs på sin spets då urvalssannolikheterna som i kommunforskningsprogrammet varierar starkt (kvoten mellan lägsta och högsta urvalssannolikhet är ungefär 1:200).
Sedan tid pågår inom statistiken en het debatt i denna fråga. Den begreppsapparat och teoribildning som finns inom den experimentella statistiken är otillräcklig för att behandla superpopulationsfallet. Den teoriutvidgning som förekommit är begreppsmässigt och innehållsmässigt komplicerad och svårtillgänglig. Ännu så länge har det inte framvuxit någon consensus beträffande frågan om man skall väga med urvalssannolikheter eller ej vid modellestimationen.

Om man väger med urvalssannolikheter riskerar man inga systematiska fel vid parameterskattningen (såvitt väntevärdesstrukturen är riktigt specificerad) och är mindre beroende av att alla modellförutsättningar är uppfyllda.

Om man inte väger med urvalssannolikheter kan man å andra sidan under vissa förutsättningar få högre precision i parameterskattningarna om alla modellantaganden är uppfyllda. Den stora risken är emellertid att urvalssannolikheterna (representerade av säg variabeln $Z_i$) är korrelerade med den beroende variabeln). Om variabeln Z inte ingår som förklaringsvariabel (men modellen ändå är sann vad beträffar väntevärdesstruktur) kan mycket allvarliga systematiska estimationsfel uppstå. Detta har bevisats för regressionsmodeller av Nathan and Holt [22].

En uppfattning som framhållits i litteraturen är att man i modellsökningsfasen av en undersökning får en bättre bild av populationens struktur om man väger med urvalssannolikheter. Därigenom skulle risken minska att man specificerar modeller som är helt inadekvata (se exempelvis Holt, Smith and Winter [12]).

Många, kanske flertalet samhällsvetenskapliga studier torde karakteriseras av just denna osäkerhet om orsakssammanhang vilken leder till att ett större antal olika modeller prövas. Personligen ansluter jag mig därför till åsikten att man bör genomföra vägningar vid estimationen.

## 1.7 Teori för analys av komplexa urval

I stort sett all teori för sambandsanalys förutsätter oberoende observationer från flerdimensionella fördelningar (i experimentella fallet ofta från olika endimensionella fördelningar). Detta är ekvivalent med obundet slumpmässigt urval från en oändligt stor population. Det avsteg från förutsättningarna som ligger i att man har ett obundet slumpmässigt urval utan återläggning (OSU) från en ändlig population har i praktiken ingen betydelse.

Om man kommer utanför denna urvalsform saknas i stort sett teori. Enstaka resultat finns för exempelvis analys av kvalitativa variabler vid stratifierat urval med OSU av element inom strata.

Grupp- och flerstegsurval uppvisar ofta utpräglad homogenitet inom grupper och delgrupper varför man i allmänhet måste

betrakta mätvärdena avseende en viss egenskap för olika element som observationer på korrelerade stokastiska variabler. (Denna korrelation är av samma art men mycket mera komplicerad än den s.k. autokorrelation eller seriekorrelation som ofta införes i modeller för tidsseriedata.) Någon teori för sambandsanalys baserade på sådana urvalsformer existerar inte bortsett från någon enstaka ytterst komplicerad ansats avseende s.k. loglinjära modeller för analys av kvalitativa data.
Några allmänna reflektioner avseende analys baserad på andra urvalsformer än slumpmässigt urval kan dock göras under förutsättning att det är relevant att beskriva sambandet i hela populationen med en enda modell.

a. Stratifierat elementurval med OSU inom strata och samma urvalsfraktion inom alla strata kan oftast betraktas som ekvivalent med OSU från hela populationen. Skillnaden ligger främst i att det stratifierade urvalet garanterar en jämnare fördelning av urvalet över populationen.

b. Självvägande urval (samma urvalssannolikhet för varje element i populationen) i form av grupp eller flerstegsurval ger "fläckvisa" observationer av populationen. De förväntas emellertid inte ge någon systematisk överrepresentation för någon del av populationen. Parameterskattningar avseende en modells väntevärdesstruktur borde därför i de flesta fall inte bli bekräftade med någon större systematisk snedvridning om analysteori för OSU användes. Däremot torde vanligen estimatorernas medelfel underskattas, i många fall förmodligen kraftigt. Konfidensintervall och signifikanstest kommer då att starkt överdriva säkerheten i slutsatserna. De reella felriskerna blir med andra ord betydligt större än de nominella.

c. Med urval som icke är självvägande (varierande urvalssannolikheter) befaras en systematisk snedvridning av parameterskattningar om standardteorin för OSU användes. (Detta torde med säkerhet vara fallet om urvals-

sannolikheten samvariera med den beroende variabeln.) Denna snedvridning kan förmodligen elimineras medelst ett vägningsförfarande baserat på inklusionssannolikheter för element. Sådant vägningsförfarande saknas ofta i standardprogrampaket. Sådana vägningar eliminerar emellertid inte de ofta ytterst allvarliga underskattningar av medelfelen som erhålles om standardprogrammen används.

Vid analys av komplexa urval (bortsett från de deskriptiva "samplingteoretiska" delarna) förbises eller nonchaleras vanligen de problem som orsakas av det komplexa urvalsschemat.

# 2. MODELLTYPER OCH METODER

## 2.1 Symmetriska och asymmetriska modeller

Det finns två huvudtyper av statistiska problemställningar vilka med avseende på de i modellen ingående variablernas roll kan kallas asymmetriska resp. symmetriska.

a. Asymmetriska
Modellen beskriver hur en responsvariabel (ev. flera responsvariabler) påverkas av ett antal förklaringsvariabler.

b. Symmetriska
Modellen beskriver hur ett antal variabler samvarierar.

Beroende på typen av problemställning och beroende på om ingående variabler är kvantitativa och/eller kvalitativa kan sakproblemet analyseras med olika multivariata modeller och metoder. Det finns ett stort antal sådana metoder. Tabell 2.1 täcker ett antal metoder där själva modellen är av primärt intresse och inte bara utgör ett medel för exempelvis en klassificering av objekt.

Tabell 2.1 Vanliga multivariata analysmetoder.

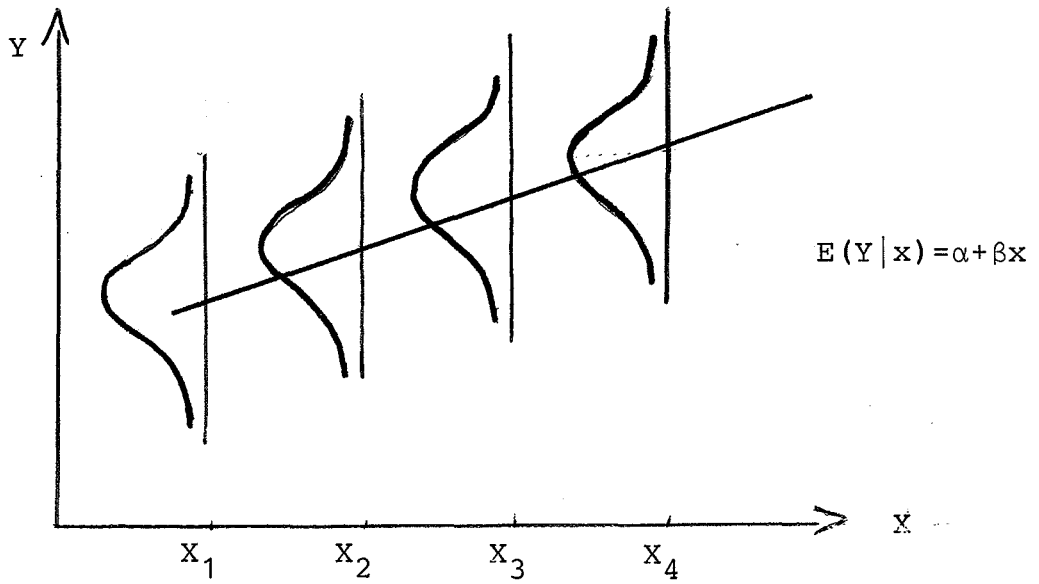
| Typ av problemställning | Responsvariabel | Förklaringsvariabler | Metod |
|---|---|---|---|
| 1. Asymmetrisk | Kvantitativ | Kvantitativa | Regressionsanalys |
| 2. Asymmetrisk | Kvantitativ | Kvantitativa och kvalitativa | Regressionsanalys med dummyvariabler Kovariansanalys |
| 3. Symmetrisk | _____ | Kvantitativa | Korrelationsanalys |
| 4. Asymmetrisk | Kvantitativa | Kvalitativa | Variansanalys (Regressionsanalys med dummyvariabler) |
| 5. Asymmetrisk | Kvalitativ | Kvantitativa | Logitanalys, Probitanalys |
| 6. Asymmetrisk | Kvalitativ | Kvalitativ | Linjära modeller, Logitanalys, Allmän loglinjär analys |
| 7. Symmetrisk | _____ | Kvalitativ | Allmän loglinjär analys |

## 2.2 Regressions- och korrelationsanalys för olika undersökformer

Regressionsanalys och korrelationsanalys används vid utvärdering av resultat från experiment, urvalsundersökningar och tidsserieundersökningar. De statistiska modeller som beskriver sannolikhetsfördelningar för variabler är emellertid av helt olika slag för de olika undersökningstyperna. Regressionsanalys är adekvat för samtliga undersökningstyper medan korrelationsanalys i huvudsak endast är meningsfull i urvalsundersökningar vilket belyses nedan.
Genomgången av de olika modelltyperna konkretiserar också den allmänna diskussionen om olika arter av modeller vid experiment och urvalsundersökningar som fördes i kapitel 1.4.

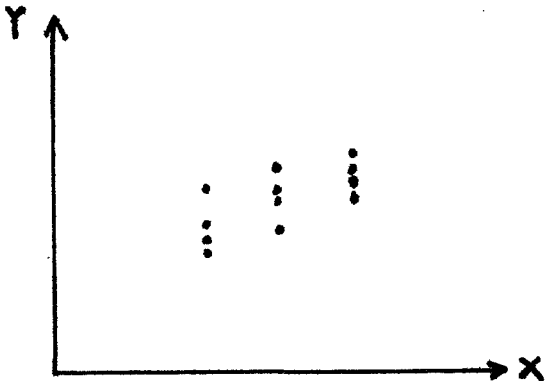
## 2.3 Experiment

För varje på förhand vald nivå för en variabel x (eller varje kombination av nivåer för ett antal variabler) göres observationer på en stokastisk variabel Y som antas följa en viss sannolikhetsfördelning. Om man gör ett antagande om hur väntevärdet för den stokastiska variabeln Y beror av x-nivåerna erhålles en regressionsmodell. Grafiskt kan modellen illustreras som i figur 2.1 där förväntat Y-värde ändras linjärt med x-värdet.

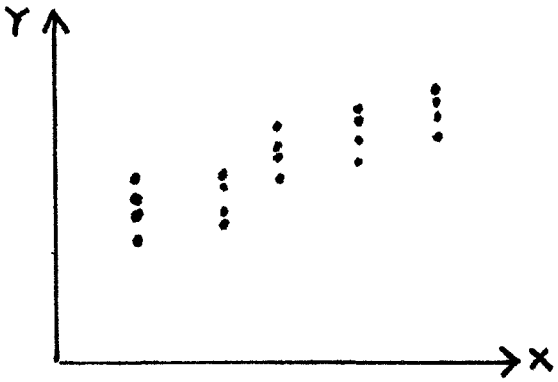



Figur 2.1 Enkel linjär regressionsmodell för experimentell undersökning.

De observationer som erhålles i experimentserien kan beskrivas enligt figur 2.1a och 2.1b.




Figur 2.1 a Punktmönster vid observationer på tre x-nivåer

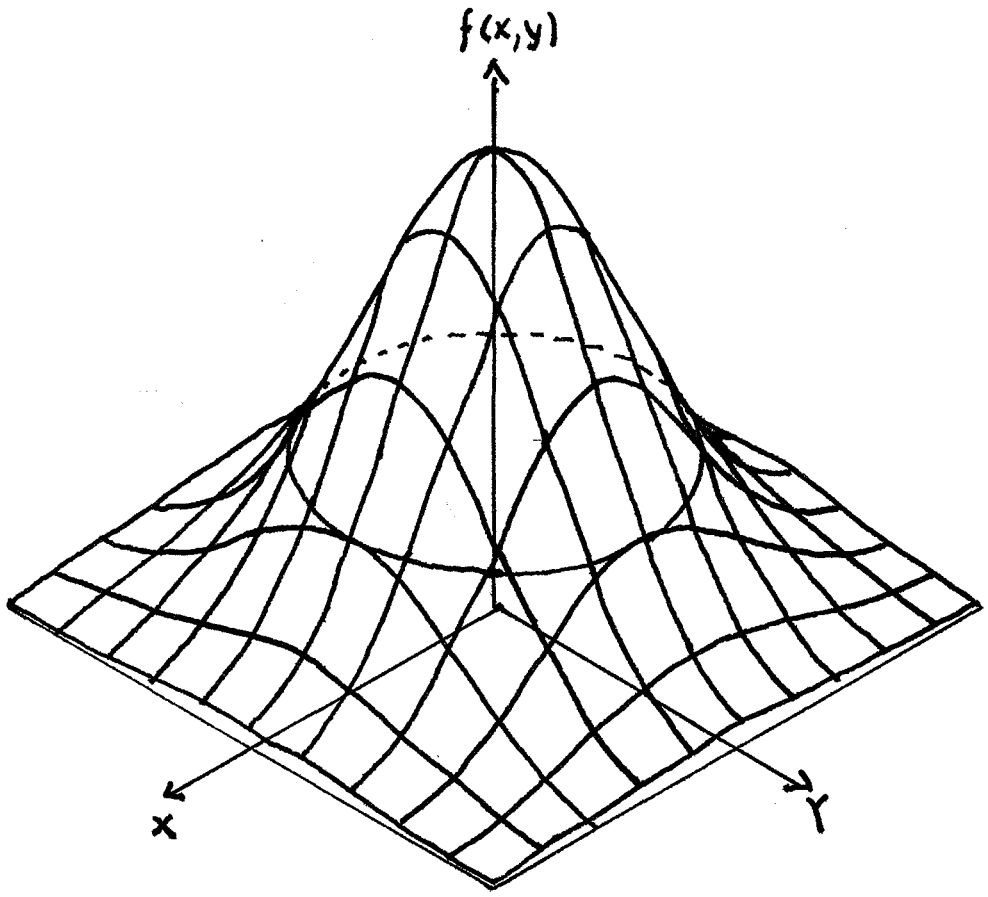



Figur 2.1 b Punktmönster vid observationer på fem x-nivåer

Punktmönstrets allmänna utseende kan påverkas genom vårt val av x-nivåer. Det är därför inte meningsfullt att beräkna något mått på graden av samvariation mellan Y och x. Uttryckt i statistisk terminologi kan man säga att korrelationer inte är definierade eftersom det endast föreligger observationer på en enda stokastisk variabel Y som observeras för vissa på förhand bestämda x-nivåer.

## 2.4 Urvalsundersökningar

Fördelningen av värden på ett antal variabler i en population av individer kan ges en förenklad beskrivning med hjälp av en sannolikhetsfördelning för en flerdimensionell stokastisk variabel. I fallet med två stokastiska variabler X och Y kan fördelningen illustreras enligt figur 2.2.




Figur 2.2 Sannolikhetsfördelning för bivariat normalfördelning.

Flerdimensionella normalfördelningar spelar en central roll då samtliga variabler som studeras är slumpmässiga. Den bivariata normalfördelningen karakteriseras entydigt av fem parametrar: väntevärde och varians för Y, väntevärde och varians för X samt korrelationskoefficienten mellan X och Y.

Då både X och Y är slumpmässiga är det alltså meningsfullt att använda urvalets korrelationskoefficient som uppskattning av populationens korrelationskoefficient.

Den som använder bivariat normalfördelning som modell för en viss företeelse har därmed förutsatt, kanske utan att tänka på det, en rad egenskaper för variablerna.

1. Marginalfördelningen för X är normal, likaså marginalfördelningen för Y.

2.a För ett givet y-värde följer X en normalfördelning.

Den betingade variansen i X är lika stor för alla y-värden.

b För ett givet x-värde följer Y en normalfördelning. Den betingade variansen i Y är lika stor för alla x-värden.

3.a Regressionen av X på Y är linjär.

b Regressionen av Y på X är linjär.

Multivariata normalfördelningar har också vissa mycket speciella egenskaper. Dessa exemplifieras här med den tredimensionella normalfördelningen.

1.a De endimensionella marginalfördelningarna för X, Y och Z är normala.

b De tvådimensionella marginalfördelningarna för (X,Y), (X,Z) och (Y,Z) är bivariat normala.

2.a För givna x- och z-värden är den betingade fördelningen för Y normal. Variansen är densamma för alla givna x- och z-värden.

b Motsvarande gäller för Z och X-variablerna.

3.a Regressionen av Y på X och Z är linjär.

b Motsvarande gäller för X och Z.

4.a Regressionen av Y på X är (bortsett från nivåkonstanten $\alpha$) densamma för alla z-nivåer.

b Motsvarande gäller för övriga betingade regressioner.

Detta är ett annat sätt att uttrycka punkt 3. Denna egenskap kan yttryckas på ännu ett sätt vilket ges i punkt 5.

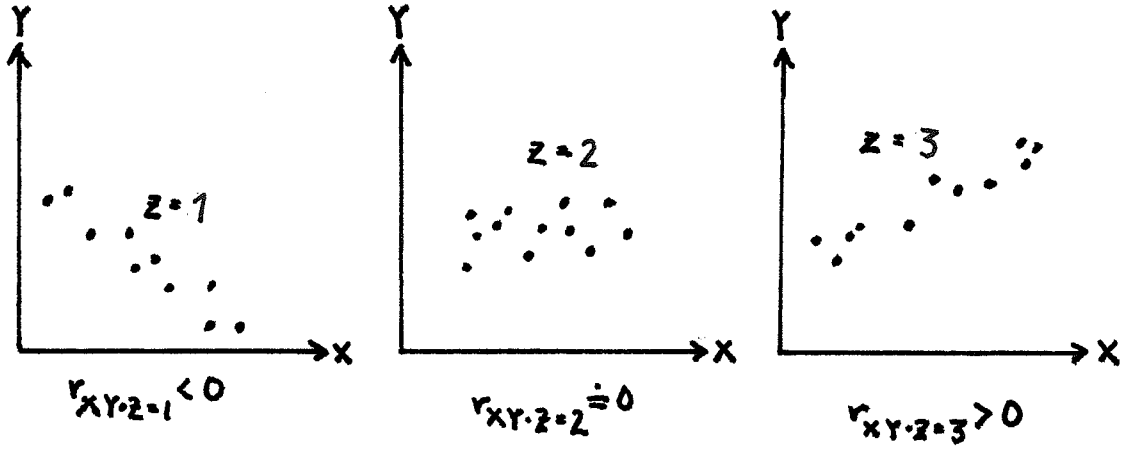
5.a Den betingade korrelationen mellan X och Y är lika stor för alla z-nivåer.

b Motsvarande gäller för övriga betingade korrelationer.

6. Detta gemensamma värde för alla betingade korrelationer mellan X och Y är lika stort som den partiella korrelationen mellan X och Y.

Vid studium av populationer som har klart olika betingade korrelationer mellan X och Y på olika Z-nivåer är det därför inte meningsfullt att beräkna den partiella korrelationen mellan X och Y givet Z, vilken kan uppfattas som

ett slags vägt medelvärde av den betingade korrelationen för olika Z-nivåer (se figur 2.4).




Figur 2.4 Betingade korrelationen mellan X och Y starkt negativ, nära 0 resp. starkt positiv för Z=1, Z=2 resp. Z=3. Den partiella korrelationen mellan X och Y givet Z blir nära 0 och är meningslös som beskrivning av samvariationen mellan X och Y då effekten av Z "elimineras".

Vid korrelationsstudier kan det vara lämpligt att klassindela datamaterialet efter Z och beräkna den partiella korrelationen mellan X och Y för varje klass. Om dessa korrelationer är någorlunda lika är det relevant att beräkna en partiell korrelation för hela materialet. I annat fall bör den partiella korrelationen redovisas klassvis.

## 2.5 Tidsserieundersökningar

Vid konstruktion av modeller för regressions- och korrelationsanalys på tidsseriedata föreligger speciella komplikationer. Det fenomen som skall studeras och det sammanhang i vilket det uppträder genomgår förmodligen gradvisa förändringar varför det är svårt att avgöra hur lång tidsserie som kan anses ge relevant underlag för analysen. I många sammanhang förekommer också politiska beslut och åtgärder som ger plötsliga förändringar i ett system. Frågan är då om man på något sätt kan representera en sådan förändring i en modell eller om detta helt invaliderar äldre data. Karakteristiskt för tidsseriemodeller är också att man i allmänhet måste förutsätta att avvikelser från väntevärdet vid en tidpunkt (eller under en tidsperiod) är beroende av avvikelser vid närmaste föregående tidpunkter (tidsperioder).
I denna rapport kommer inte de speciella problem som är förknippade med tidsserieundersökningar att behandlas.

# 3. ESTIMATIONSMETODER

Vissa regressionsansatser fungerar inte i speciella situationer vilket diskuteras i senare kapitel. För att förstå vissa delar av resonemanget i dessa måste man känna till några grundbegrepp inom estimationsteorin. Därför ges en kort översikt av dessa grundbegrepp nedan.

## 3.1 Goda estimationsegenskaper

Vid skattning av en parameter i en modell är det uttryckt i allmänna ordalag önskvärt att totala felet, som utgöres av summan av systematiskt och slumpmässigt fel, minimeras. Om möjligt vill man naturligtvis använda en estimator fri från systematiskt fel. I statistisk teori formaliseras dessa goda egenskaper till

1. väntevärdesriktighet. För given stickprovsstorlek blir förväntat värde för estimatorn det sanna parametervärdet.
2. minimal varians. Förutsatt att väntevärdesriktighet föreligger är estimatorn med lägsta variansen bäst.

Om en estimator inte är väntevärdesriktig kan den ändå vara bra om det systematiska felet är litet. Man kräver av sådana estimatorer också att det systematiska felet skall minska med ökande stickprovsstorlek. Detta allmänna önskemål formaliseras till

3. konsistens. En estimator är konsistent om det systematiska felet (kallat "bias" och definierat som skillnaden mellan estimatorns förväntade värde och sanna parametervärdet) i viss mening konvergerar mot 0 då urvalsstorleken växer mot oändligheten (resp. växer mot populationsstorleken vid urval utan återläggning från ändlig population). Det aktuella konvergensbegreppet är komplicerat och definieras ej här.

## 3.2 Estimationsmetoder

Det förekommer olika estimationsprinciper inom statistiken vilka leder till olika estimationsmetoder. De generellt

sett viktigaste bland dessa är minsta-kvadratmetoden (MK-metoden) och maximumlikelihoodmetoden (ML-metoden). Den förra kräver inget fördelningsantagande om slumpkomponenten i motsats till den senare. Båda metoderna garanterar optimala estimatorsegenskaper under vissa förutsättningar. Om dessa förutsättningar inte är uppfyllda kan en skattningsmetod som fungerar bra för en besläktad modell medföra allvarliga systematiska skattningsfel för den aktuella modellen.

Det är därför viktigt att vid analys av empiriska undersökningar dels kunna specificera en utfallsmodell dels kunna avgöra om förutsättningarna föreligger för att en viss estimationsmetod skall ge adekvata skattningar. Minsta-kvadratmetoden som är en bra metod allmänt sett ger till exempel systematiska skattningsfel för speciella regressionsmodeller (vissa flerekvationsmodeller).

a. Minsta-kvadratmetoden. Denna innebär i princip att man bestämmer parameterskattningen så att summan av kvadrerade avvikelser mellan Y-observationerna och väntevärdet för Y minimeras. (En allmän och stringent definition återfinnes exempelvis i Kendall and Stuart [16 ]).

Vid parameterskattning i regressionsmodellen $E(Y_i) = \alpha+\beta x_i$ väljes således skattningar $\hat{\alpha}$ och $\hat{\beta}$ så att kvadratavvikelsesumman

$$Q=\Sigma(y_i-(\hat{\alpha}+\hat{\beta}x_i))^2$$

minimeras. Denna grundform av minsta-kvadratmetoden i vilken alla observationer ges samma vikt (man talar också om ovägd minsta-kvadratskattning) benämns i engelspråkig litteratur the Ordinary Least Squares Method och anges mycket ofta endast som "the OLS Method". Det förekommer även olika vägda minsta-kvadratmetoder, en enklare "Weighted Least Squares Method" (WLS) i vilken vikterna bestäms av observationernas varianser och en "Generalized Least Squares Method (GLS) som tar hänsyn både till varianser för observationer och kovarianser mellan observationer (Se läroböcker i regressions-

analys, såsom Draper and Smith [ 3 ] eller Hanushek and Jackson [11 ]).

För en vid klass av modeller kallade den linjära modellen för vilken observationerna

- har ett väntevärde som är linjärt i vissa parametrar (och kan bero på variabler $x_1,\ldots,x_k$)
- har konstant varians
- är parvis okorrelerade

har minstakvadratmetoden (OLS) en optimal egenskap. Om man håller sig till klassen av estimatorer som är linjära i observationerna ger OLS till varje parameter en estimator som är väntevärdesriktig och har minimal varians inom klassen.

b. Maximumlikelihoodmetoden (ML-metoden).

Likelihoodfunktionen för en diskret variabel anger sannolikheten att erhålla det givna samplet. (Definitionen för en kontinuerlig variabel är analog men likelihooden kan ej tolkas som en sannolikhet.) Denna sannolikhet beror av vissa parametrar. ML-metoden innebär att man som parameterestimat väljer de värden som bäst "förklarar" det givna samplet. ML-estimaten är därför de värden på parametrarna som maximerar likelihoodfunktionen.

De goda egenskaper hos ML-estimatorerna som motiverar ML-metodens användning är av komplicerad natur och kan inte här ges en strikt definition. (Se exempelvis Kendall and Stuart vol. 2 [16 ]).

En av dessa egenskaper är att om det finns en väntevärdesriktig estimator som uppnår en teoretisk minsta möjliga varians (Cramér-Raogränsen) så erhålles denna estimator med ML-metoden.

En annan egenskap är att under vissa milda regularitetsvillkor är ML-estimatorer konsistenta, har asymptotiskt (dvs. då urvalsstorleken växer mot oändligheten) den teoretiskt minsta möjliga variansen och är asymptotiskt normalfördelade.

Det viktigaste motivet till att använda ML-metoden är således att under allmänna och milda förutsättningar garanterar ML-metoden goda storsampleegenskaper. Vad som skall betraktas som ett stort sample beror emellertid på fördelningstyp för observationerna och på parametertyp.

En jämförelse av MK-metoden och ML-metoden visar att den förra under speciella förutsättningar garanterar goda skattningsegenskaper för varje given urvalsstorlek. Den senare garanterar under mycket allmänna förutsättningar också goda egenskaper men dessa gäller i gengäld endast asymptotiskt. Det är också värt att ånyo notera att MK-metoden inte kräver fördelningsantagande om observationerna vilket däremot ML-metoden gör. I många situationer ger MK-metoden och ML-metoden samma estimatorer.

# 4. SPECIFIKATION AV REGRESSIONSMODELLER

I detta kapitel samt i kapitlen 5 och 6 diskuteras specifikation av regressionsmodeller. Denna modelltyp är central i surveys. Diskussionen illustrerar också väl det allmänna tillvägagångssättet och problemen vid modellspecifikation och modellbyggnad.
Hanushek and Jackson säger i kap 4 av Statistical Methods for Social Scientists [11 ]

"Certainly it is easy to program a computer to calculate the coefficient estimates and the summary statistics from the preceding chapter. In fact, computer programs to do just that are plentiful. *But the actual application of ordinary least squares is not just a technician's job.* First, the construction and interpretation of the actual model demands extensive knowledge of the relevant social science theories and of previous empirical work. Considerable expertise and experience are required to decide which variables should be included in any analysis. This topic, *model specification*, is the subjekt of section 4.3 and is possibly the most important topic in this book".

Sann modell. Vid konstruktion av en modell av en viss företeelse måste det alltid ske en avvägning mellan enkelhet och modellens förmåga att ge en realistisk bild av företeelsen ifråga. Oavsett hur mycket kunskap som funnes om företeelsen skulle det ändå alltid bli en bedömningsfråga hur en adekvat modell skulle se ut. För att man teoretiskt skall kunna jämföra modeller och härleda estimatorers egenskaper krävs det som en referensbas ändå föreställningen att det existerar en sann modell.

## 4.1 Modellspecifikation

En modellspecifikation kan göras mer eller mindre speciell. Ibland specificeras sannolikhetsfördelningen exakt så när som på värdet på en eller flera okända parametrar. I andra fall undviker man att ange fördelningstypen för slumptermerna. Nedan demonstreras modellspecifikationen för ett enkelt experiment. Man slipper då ifrån alla modellproblem som är förknippade med surveys. De senare behandlas i kap. 6.

Exempel. Ett nytt blodtryckssänkande medel skall studeras i ett experiment med en grupp försökspersoner. Dessa personer har i utgångsläget samma blodtryck och utgör även i övrigt en homogen grupp. Personerna ges olika doser av det aktuella medlet. Fördelning av doser till patienterna sker fullständigt randomiserat (med hjälp av slumptalstabell).

Det är naturligt att beskriva storleken av blodtrycks-sänkningen med hjälp av en regressionsmodell. Sänkningen Y antas bestå av en systematisk del vars storlek bestäms av dosen x av medlet samt en stokastisk del $\varepsilon$ (positiv eller negativ) som bestäms av slumpen.

## 4.2 Egenskaper som specificeras i regressionsmodeller

I modellspecifikationen beskrivs dels det systematiska sambandet dels slumpvariationen. Följande punkter ingår vid specifikation av en regressionsmodell.

a. På vilket sätt förväntat Y-värde förändras med x-värdet.

b. Hur spridningen för Y-variabeln, eller ekvivalent för slumpkomponenten, varierar med x-värdet.

c. Huruvida observationerna på Y-variabeln (ekvivalent

slumptermerna) är oberoende av varandra eller ej. (Ofta nöjer man sig med att ange korrelationsegenskaper vilket ur skattningssynpunkt kan betraktas som ekvivalent om man använder minsta-kvadratmetoden.)

d. Fördelningsform för Y-variabeln (ekvivalent slumpkomponenten) för ett givet x-värde.

Punkterna a-c bestämmer en klass av modeller. I många fall begränsar man sig till en enda modelltyp genom att i en fjärde punkt d även antaga en speciell fördelningsform för Y-variabeln för givet x-värde, vanligen normalfördelning.

## 4.3 Specifikation av standardmodellen för enkel linjär regression

Utfallet vid i-te observationen på en responsvariabel kan anges som ett väntevärde $\mu_i$ samt en additiv avvikelse $\varepsilon_i$ från detta väntevärde.

$$Y_i = \mu_i + \varepsilon_i$$

I blodtrycksexemplet ovan kan det kanske vara rimligt att anta ett linjärt samband mellan förväntad blodtryckssänkning $\mu_i$ och dosen $x_i$ förutsatt att dosen ligger mellan vissa gränser. Vi antar således att sambandet mellan förväntad sänkning $\mu_i$ och dos $x_i$ kan beskrivas med en rät linje

$$\mu_i = \alpha + \beta x_i$$

Den allra enklaste regressionsmodellen (standardmodellen) för linjär regression kan specificeras på följande sätt

a. $Y_i = \alpha + \beta x_i + \varepsilon_i$

med

$E(\varepsilon_i) = 0$

b. Var $(\varepsilon_i) = \sigma^2$

Variansen för $\varepsilon_i$ (och därmed för $Y_i$) antas vara lika stor för alla x-värden.

c. Kov $(\varepsilon_i, \varepsilon_j) = 0$ för $i \neq j$.

Olika residualer antas vara okorrelerade.

(d). $\varepsilon_i$ är normalfördelat.

Om specifikationen endast omfattar punkterna a-c talar man om "de svaga antagandena". Tillägg av punkt d ger "de starka antagandena".

## 4.4 Standardmodellen för multipel linjär regression

Om väntevärdet för Y förändras linjärt med regressorerna $X_1, \ldots, X_k$ kan det i:te utfallet betraktas som en observation på den stokastiska variabeln

$$Y_i = \alpha + \beta_1 X_1 + \ldots + \beta_k X_k + \varepsilon_i$$

I det enklaste fallet antas slumptermen $\varepsilon_i$ ha samma egenskaper som i kap. 4.3.

Direkt och total effekt. Om variabel $X_j$ ökas en enhet medan övriga X-variabler förblir oförändrade förväntas en förändring av Y med $\beta_j$ enheter. Detta kallas ofta den direkta effekten av $X_j$ på Y. I ett kausalt system medför en förändring av $X_j$ ibland förändringar av andra X-variabler vilka i sin tur påverkar Y. Summan av $X_j$:s direkta effekt på Y och $X_j$:s indirekta effekter på Y via andra X-variabler kallas $X_j$:s totala effekt på Y.

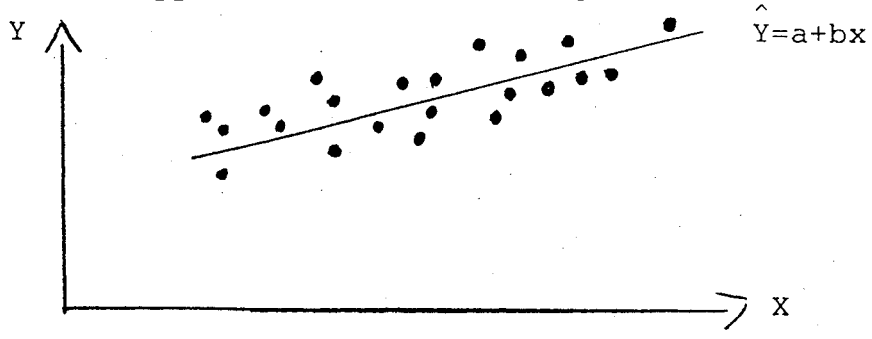
Definitionen förutsätter att Y ligger sist i ett kausalt system och inte i sin tur påverkar någon eller några X-variabler. I pathanalysen beskrivs sådana kausala kedjor och effekterna kan studeras med hjälp av regressionsanalys. (För en beskrivning se exempelvis Kendall and O´Muircheartaigh: "Path Analysis and Model Building" [15] som ingår i World Festility Surveys (WFS) serie av beskrivningar av statistiska metoder och ger illustrationer med hjälp av data från WFS).

# 5. FELSPECIFIKATION OCH MODELLALTERNATIV

## 5.1 Metoder att upptäcka fel i modellspecifikationen.

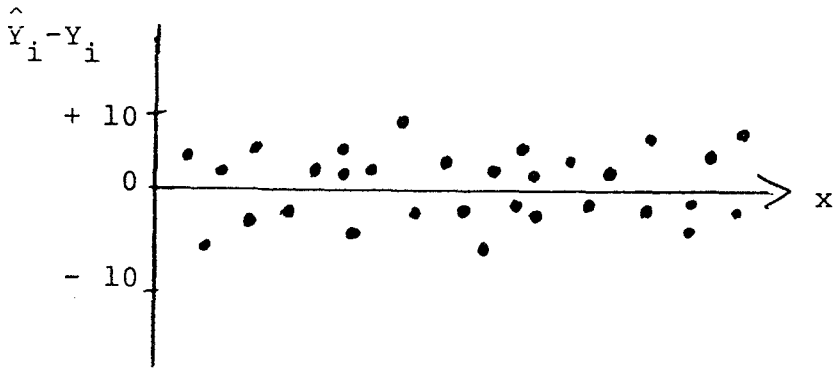
Statistiska hjälpmedel för att studera huruvida en viss modell ger en adekvat beskrivning av den studerade företeelsen består av olika former av residualplottar, av vissa enkla mått (determination, reststandardavvikelse, Mallows' $C_p$) samt av statistiska test. I fråga om linjära regressionsmodeller finns metodiken beskriven i handböcker såsom Draper and Smith, Applied Regression Analysis, 2 nd ed. [ 3 ] kapitlen 3 och 6 (kapitel 8 ger en mera allmän beskrivning av modellbyggnadsprocessen i multipel regressionsanalys) och den enklare Younger, A Handbook for Linear Regression [ 31] kapitlen 9, 14 och 15.2. Grundläggande för modellspecifikationen måste emellertid alltid vara existerande teoribildning inom området och tidigare empiriska studier.

Nedan ges en kort beskrivning av residualplottarna.
I enklaste fallen med standardmodellen för enkel linjär regression bör spridningsdiagrammet då den beroende variabeln Y plottas mot förklaringsvariabeln X ha ett utseende av den typ som förekommer i figur 5.1.




Figur 5.1 Normal spridningsbild då observationerna är genererade enligt standardmodellen för enkel linjär regression.

Om man i stället plottar residualerna $Y_i - \hat{Y}_i$ från regressionslinjen mot x bör standardmodellen ge upphov till en spridningsbild enligt figur 5.2.

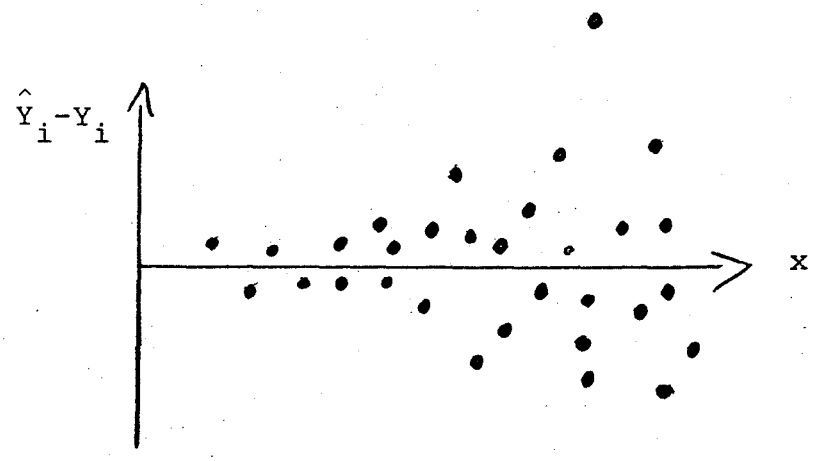



Figur 5.2 Residualdiagram. Normal spridningsbild då observationerna är genererade enligt standardmodellen för enkel linjär regression.

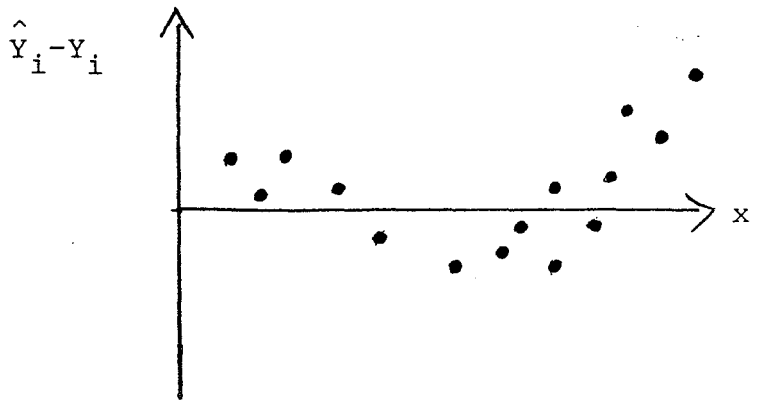
Vid enkel linjär regression ger plottning av Y mot x och plottning av residualer mot x ekvivalent information. Vid multipel regression är det däremot endast meningsfullt med residualplottar.

Figurerna 5.3 a-d visar mönster som tyder på att olika typer av avvikelser från standardmodellen kan föreligga.

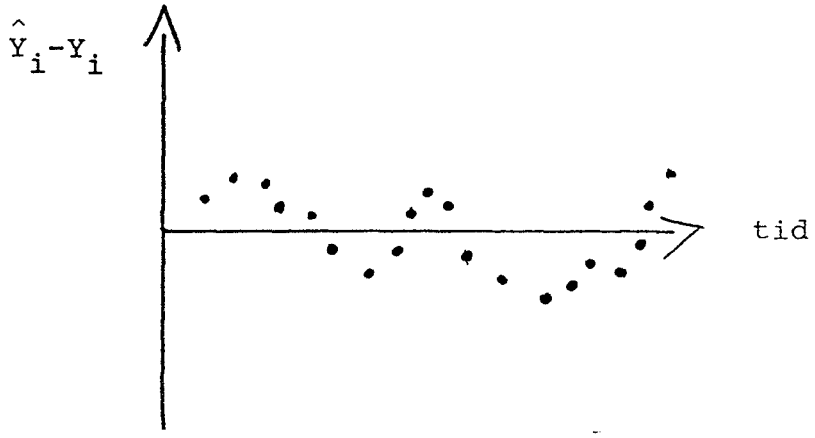
a.



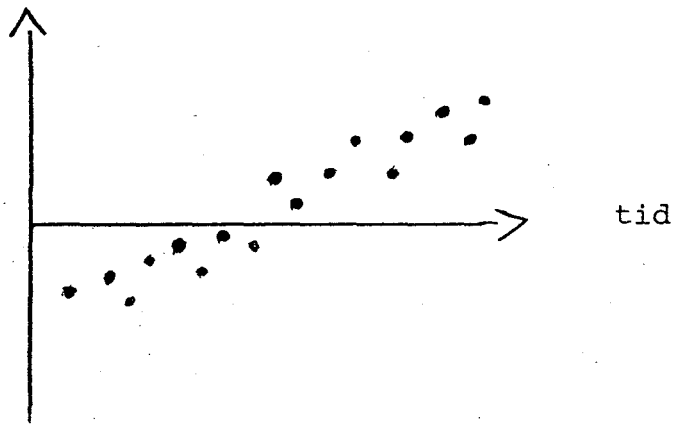
b.




c.




d.




Figur 5.3 Residualmönster vid skattning av standard-modellen.

I samtliga fall kan naturligtvis modellen vara ofullständig och bör då kompletteras med en eller flera förklaringsvariabler. Om detta inte är fallet tyder figur 5.3 a på att den sanna variansen för Y ökar med stigande x-värde.
Ett punktmönster enligt figur 5.3 b utgör en indikation på att modellen bör göras icke-linjär.
Om residualplotten, som i figur 5.3 c, visar att i tiden på varandra följande residualer tenderar att ha samma tecken är troligen en regressionsmodell med autokorrelerade feltermer adekvat.
I figur 5.3 d är residualerna starkt korrelerade med tiden vilket tyder på att ytterligare en eller flera variabler bör införas i modellen.

## 5.2 Effekter av specifikationsfel

Specifikationsfel får olika konsekvenser beroende på vari felet består. Det uppstår i olika situationer

- systematiska fel av skilda slag vid skattning av variabeleffekten
- ökade slumpfel för parameterskattningar
- systematisk underskattning av felmarginaler för estimatorer

Avvikelser från standardantagandena a-e behandlas punkt för punkt nedan.

## 5.3 Väntevärdesantagande - felaktigt utelämnade variabler

Om Y-variabeln beror även av variabler som ej ingår i modellen erhålles i allmänhet systematisk felskattning av effekten av ingående variabler. Ett enkelt fall kan belysa förhållandet.

Den sanna modellen är

$$Y = \alpha+\beta_1 x_1+\beta_2 x_2+\varepsilon$$

men man skattar en enkel linjär regression av Y på enbart $x_1$, d.v.s.

$$\hat{Y} = a+b_1 x_1$$

Då blir förväntat värde av b,

$$E(b_1) = \beta_1+\beta_2 b_0$$

där $b_0$ är stickprovets regressionskoefficient vid den enkla linjära regressionen av $x_1$ på $x_2$. (Se exempelvis Snedecor and Cochran [25]).
Förutsatt att $x_1$ och $x_2$ är korrelerade, vilket alltid är fallet i icke-experimentella undersökningar, erhålles således ett systematiskt fel $\beta_2 b_0$ vid skattning av $\beta_1$.
En utelämnad förklaringsvariabel kan leda både till teckenfel vid parameterskattning och till ett systematiskt fel som är långt större än sanna parametervärdet. Trots detta kan ofta den skattade standardavvikelsen (medelfelet) för koefficientskattningen vara liten och därmed antyda ett säkert resultat. I många fall ger utelämnande av relevanta variabler inga spår i residualplottar i form av stor spridning eller avvikande mönster.
En felspecificerad modell som inkluderar relevanta variabler men saknar vissa icke-linjära termer är däremot lättare att avslöja i residualplottar, som i sådana fall i allmänhet uppvisar icke-linjära mönster.

## 5.4 Väntevärdesantagande - felaktigt medtagna variabler

Variabler som inte påverkar Y men ändå felaktigt medtagits i modellen orsakar inget systematiskt fel vid skattning av regressionskoefficienten för variabler som verkligen påverkar Y. Däremot ökas variansen för dessa koefficientskattningar. Ökningen kan bli kraftig om den "falska" regressorn har en stark multipel korrelation med "sanna" regressorer.

## 5.5 Icke konstant varians

Situationer då variansen i Y-variabeln ökar med stigande x-värden är vanliga. Om detta är enda avvikelsen från standardmodellen för enkel linjär regression leder standardestimation av regressionskoefficienten inte till något systematiskt fel. Man bör emellertid överväga att övergå till vägd regression om variansen i Y ökar mycket starkt med stigande x-värde. Med lämplig viktuppsättning erhålles då lägre medelfel för estimatorn av $\beta$. (Se lärobok i regressionsanalys, exempelvis Draper and Smith [3 ] eller Younger [31]).

## 5.6 Korrelerade residualer

Vid tidsserieundersökningar är det ofta adekvat att ansätta en modell med korrelation mellan i tiden på varandra följande residualer. Om sanna regressionssambandet är linjärt förväntas standardestimation i en sådan situation ge underskattning av variansen i skattningen av regressionskoefficienten $\beta$. Effekten blir då i allmänhet för korta konfidensintervall och vid test stor risk för falska signifikanser. En lösning kan vara att övergå till estimation baserad på differenser mellan i tiden på varandra följande variabelvärden $Y_t - Y_{t-1}$ respektive $x_t - x_{t-1}$, där t anger tidpunkt eller tidsperiod. Det är dock en besvärlig avvägningsfråga att avgöra när detta är fördelaktigt. (Se lärobok i ekonometri eller regressionsanalys, exempelvis Wonnacott and Wonnacott [30] eller Younger [31]).

## 5.7 Normalitetsantagandet

Statistisk teori är till största delen utvecklad under normalitetsantagande. I regressionsmodeller antas ofta den beroende variabeln $Y_i$ eller ekvivalent testtermen $\varepsilon_i$ vara normalfördelad för givet x-värde. Detta leder till att den exakta fördelningen för en rad olika estimatorer och testvariabler är kända (t-, $X^2$- eller F-fördelningar). Normalitetsegenskapen ägnas emellertid ofta alltför stor uppmärksamhet på bekostnad av andra modellegenskaper som vanligen är viktigare. De statistiska skattnings- och testprocedurerna är nämligen ganska robusta gentemot avvikelser från normalitet förutsatt att urvalet är stort. Procedurerna fungerar relativt tillfredsställande även vid klara avvikelser från normalitet för Y-variabeln så länge Y-variabeln är kontinuerlig eller i varje fall kvantitativ. Resultaten bör dock tolkas konservativt om Y har en markerat sned fördelning för givet x-värde.

Det finns emellertid en situation då skattnings- och testprocedurerna fungerar mindre bra, nämligen då Y-variabeln är dikotom (0-1 variabel). I vissa sådana fall är en linjär modell helt inadekvat. I dessa fall kan det vara aktuellt att ansätta helt andra modelltyper: logitmodeller, probitmodeller eller loglinjära modeller. Problemet behandlas i ett senare kapitel.

## 5.8 Korrelation mellan x och residual

Hittills har förutsatts att en enekvationsmodell av typen

$$Y = \alpha+\beta x+\varepsilon$$

har beskrivit det aktuella problemet på ett adekvat och uttömmande sätt. Detta är emellertid inte alltid fallet. Ibland kan det behövas ett ekvationssystem för en sådan beskrivning. Inom ekonometrin har man länge använt sådana ekvationssystem för att beskriva nationalekonomisk teori. En överförenklad maroekonomisk modell brukar användas för att introducera modelltypen.

Betrakta en konsumtionsfunktion där konsumtionen (KONS) är linjär i inkomsten (INK). Med tidsseriedata och index t för att beteckna tidsperiod erhålles:

$$KONS_t = \alpha+\beta\cdot INK_t+\varepsilon_t$$

Man antar vidare att nationalinkomsten går till konsumtion och investeringar (INV). Något sparande antas ej förekomma. Det behövs således en andra ekvation för att systembeskrivningen skall bli fullständig.

Således blir den kompletta modellen

$$KONS_t = \alpha+\beta\cdot INK_t+\varepsilon \qquad (1)$$

$$INK_t = KONS_t+INV_t \qquad (2)$$

Man brukar betrakta $INV_t$ som en exogen variabel som för varje given tidpunkt t är bestämd av yttre faktorer som ej ingår i modellen.

$KONS_t$ och $INK_t$ är däremot båda endogena variabler som bestäms i modellen.

Wonnacott and Wonnacott [30] behandlar modellen och förklarar varför det uppstår korrelation mellan $INK_t$ och residual i stickprovet. Denna korrelation ger upphov till systematiska skattningsfel om enbart ekvation (1) skattas med standardmetodik (minstakvadratmetoden).

Konstruktion av modeller i form av simultana ekvationssystem och skattning av sådana modeller kräver gedigna kunskaper både om sakområdet och i regressionsanalys.

Modelltypen behandlas allmänt i läroböcker i ekonometri såsom Kmenta [18], Intriligator [14] och Wonnacott and Wonnacott [30] men även i vissa allmänt samhällsstatistiskt inriktade verk som Hanushek and Jackson [11].

## 5.9 Mätfel

En faktor som inte har med själva modellen att göra men ändå har stor betydelse för estimationsegenskaperna är förekomsten av mätfel vid observation av variablerna. Att systematiska mätfel leder till systematiska skattningsfel är självklart. Även oberoende slumpmässiga mätfel med väntevärde 0 kan emellertid ofta ge upphov till allvarliga systematiska skattningsfel. Således underskattas systematiskt lutningen vid enkel linjär regression om förklaringsvariabeln x är behäftad med slumpmässiga mätfel. Detta innebär inte att en observerad regressionskoefficient med nödvändighet har för lågt värde utan att väntevärdet av regressionskoefficienten är för lågt.

## 5.10 Interaktion

I standardmodellen för multipel linjär regression är effekten av en regressor oberoende av nivån på övriga regressorer. Låt exempelvis

$$E(Y) = \alpha+\beta_1 K_1+\beta_2 X \qquad (5.5)$$

där Y och $X_2$ är kontinuerliga och K är en indikatorvariabel som anger kön

$$K = \begin{cases} 1 & \text{för kvinnor} \\ 0 & \text{för män} \end{cases}$$

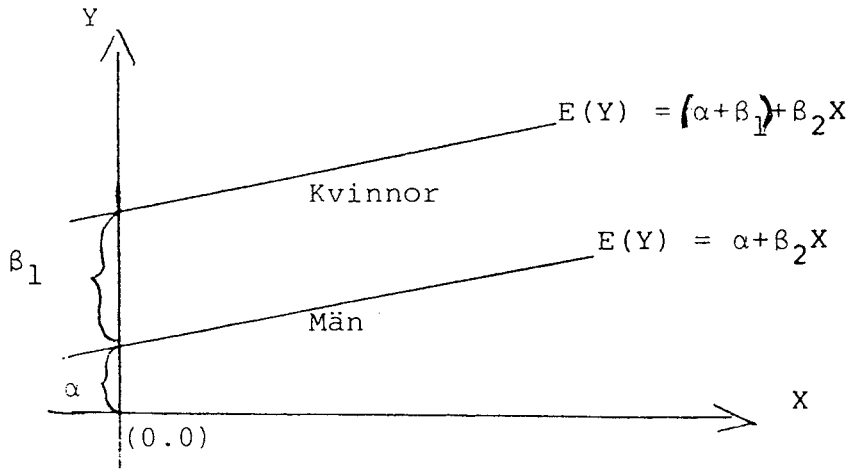
För män (K=0) faller andra termen i högerledet bort vilket ger den räta linjen

$$E(Y) = \alpha+\beta_2 X \qquad (5.6)$$

För kvinnor (K=1) blir andra termen $\beta$ vilket ger den räta linjen

$$E(Y) = (\alpha+\beta_1)+\beta_2 X \qquad (5.7)$$

Linjerna för män och kvinnor har samma lutning, angiven av regressionskoefficienten $\beta_2$, medan nivåerna skiljer sig åt. Regressionskoefficienten $\beta_1$ i (5.5) anger således nivåskillnaden (skillnaden i intercept) mellan linjen (5.6) för män och linjen (5.7) för kvinnor. Om $\beta_1>0$ illustreras förhållandet av figur 5.4. Man säger att ekvationen (5.5) och figuren 5.4 representerar additiva effekter.




Figur 5.4. Regressorerna X och K har additiva effekter.

I många situationer är inte antagandet om additivitet realistiskt. Om effekten av regressorn X är olika stark för män och kvinnor (d.v.s. för olika nivåer av variabeln K, uttryckt i allmänna termer) men sambandet mellan Y och X fortfarande är linjärt kan modellen skrivas

$$E(Y) = \alpha+\beta_1 K+\beta_2 X+\beta_3 KX \qquad (5.8)$$

För K=0 (män) försvinner andra och fjärde termen vilket fortfarande ger den räta linjen (5.6) d.v.s.

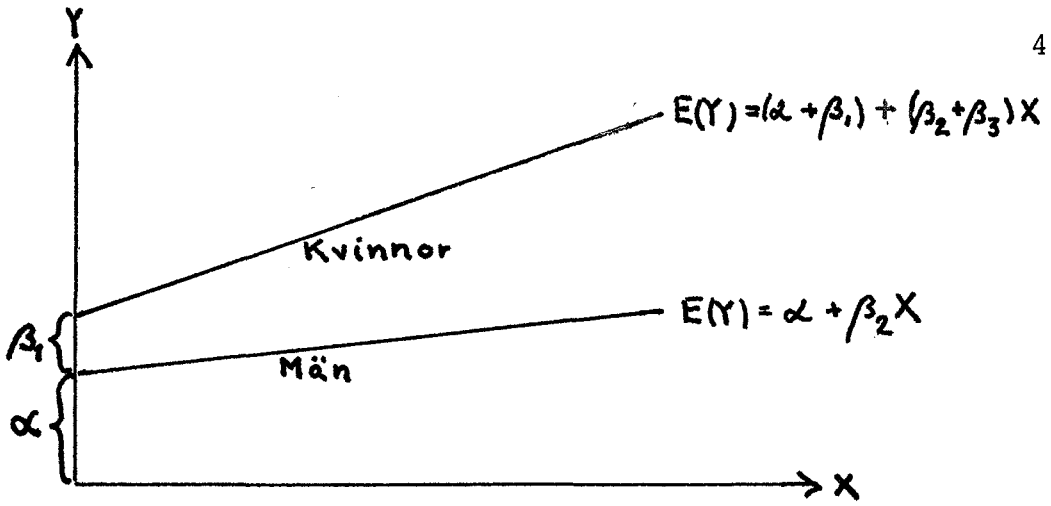
$$E(Y) = \alpha+\beta_2 X \qquad (5.9)$$

För K=1 (kvinnor) blir andra termen $\beta_1$ och fjärde termen $\beta_3 X$. Man får alltså

$$E(Y) = \alpha+\beta_1+\beta_2 X+\beta_3 X = (\alpha+\beta_1)+(\beta_2+\beta_3)X$$

En jämförelse av ekvationerna 5.9 och (5.10) för män respektive kvinnor visar att $\beta_1$ som i föregående fall anger skillnaden i intercept mellan män och kvinnor medan $\beta_3$ anger lutningsskillnaden mellan linjerna för män och kvinnor.

Om $\beta_1>0$ och $\beta_3>0$ illustreras förhållandet av principskissen i figur 5.5.




Figur 5.5. Samspelseffekt mellan K och X.

KX-termen i modell (5.8) kallas interaktionsterm eller samspelsterm. Variablerna K och X har med andra ord inte additiva effekter på Y utan samspelar i sin inverkan på Y.
Formen (5.8) kan användas även då alla variabler är kategoriska.

I kapitel 8 användes mera utvecklade former av nedanstående modelltyp. Låt A vara en indikatorvariabel som för en person anger förekomst (A=1) eller avsaknad av (A=0) viss form av politisk aktivitet. Denna förklaras av UTB (med värde 0 för obligatorisk utbildning och värde 1 för utbildning därutöver) och POL (med värde 1 om någon person i föräldrahemmet var politiskt engagerad och 0 eljest).
Både utbildning utöver den obligatoriska och politiskt engagemang i föräldrahemmet antas ha positiv effekt på sannolikheten för politisk aktivitet. Låt $\beta_u$ ange utbildningseffekt för personer som växt upp i föräldrahem utan politiskt engagemang och $\beta_p$ ange effekten för personer som endast har obligatorisk utbildning men växt upp i ett hem där någon var politiskt engagerad. För personer som vuxit upp i ett hem med politiskt engagemang och som har utbildning utöver den obligatoriska är förmodligen den kombinerade effekten av dessa båda faktorer mindre än summan $\beta_{UTB}+\beta_{POL}$ .
Man kan då ansätta modellen

$$E(A) = \alpha+\beta_u UTB+\beta_p POL+\beta_{up} UTB\cdot POL \qquad (5.11)$$

Sätter man in aktuellt värde på UTB (0 eller 1) och POL (0 eller 1) erhålles sannolikheten för politisk aktivitet (eftersom E(A) = P(A=1)) för var och en av de 4 möjliga grupperna.

UTB=0, POL=0 ger $E(A) = \alpha$

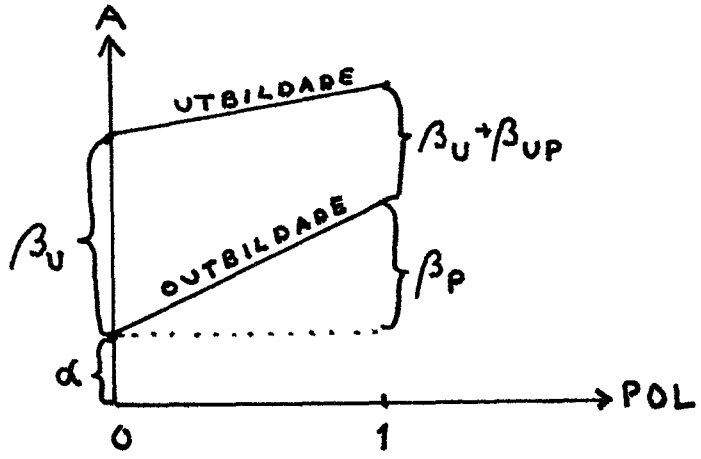
UTB=1, POL=0 ger $A(A) = \alpha+\beta_u$

UTB=0, POL=1 ger $E(A) = \alpha+\beta_p$

UTB=1, POL=1 ger $E(A) = \alpha+\beta_u+\beta_p+\beta_{up}$

Enligt förutsättningarna ovan skulle $\beta_{up}$ vara negativ.

Figur 5.6 illustrerar förhållandet.




Figur 5.6. Negativ samspelseffekt mellan kategoriska variablerna UTB och POL.

I och för sig är det möjligt att tolka termen $\beta_{up}$. I detta fall anger $\beta_{up}$ (som var negativ) hur mycket mindre utbildningseffekten var bland personer som vuxit upp i familj med politiskt engagemang än bland personer som vuxit upp i familj utan sådant engagemang. Ekvivalent anger $\beta_{up}$ hur mycket mindre effekten av politiskt engagemang var bland utbildade än bland outbildade personer. Det förefaller emellertid bättre att inte tolka termen $\beta_{up}$ separat utan att tolka följande parametrar

$\alpha$: sannolikheten för politisk aktivitet för en person utan utbildning och utan politiskt engagemang i föräldrahemmet

$\beta_u$: utbildningseffekten på sannolikheten för politisk aktivitet för en person utan politiskt engagemang i föräldrahemmet

$\beta_p$: effekten av politiskt engagemang i föräldrahemmet på sannolikheten för politisk aktivitet för en person utan utbildning

$\beta_u+\beta_p+\beta_{up}$: den kombinerade effekten av utbildning och politiskt engagemang i föräldrahemmet på sannolikheten för politisk aktivitet.

# 6. BYGGNAD OCH TOLKNING AV REGRESSIONSMODELLER

## 6.1 Förklaringsmodeller och prediktionsmodeller

Regressionsanalysen används ofta för att förklara hur en beroende variabel Y påverkas av ett antal regressorer $X_1, X_2, \ldots, X_k$. I vissa fall krävs komplexa modeller i form av simultana ekvationssystem för en sådan förklaring. Bl.a. ekonomer utnyttjar regressionsanalys för att konstruera prognosmodeller. Härvid är det naturligtvis en styrka om man kan finna en förklaringsmodell men ofta måste man nöja sig med en modellbeskrivning av en stabil samvariationsstruktur mellan Y och andra på förhand kända eller lätt prognostiserbara variabler $X_1, X_2, \ldots, X_k$. Vid konstruktion av förklaringsmodeller kommer själva modellens utseende och regressionskoefficienternas storlek i fokus. Prognosarbete är däremot inriktat mot att finna en modell som ger små prognosfel (skillnad mellan utfallet och prognosen enligt modellen för viss tidpunkt). Modellens exakta utseende och regressionskoefficienter är då av mera underordnat intresse. Ibland kan det finnas flera sinsemellan ganska olika men i stort sett likvärda prognosmodeller.

Modellproblem i surveys. Inom samhällsvetenskaperna saknas oftast etablerad teori som tillåter att man konstruerar förklaringsmodeller för olika sakförhållanden. Vid surveys måste man därför i många fall nöja sig med att kontruera prediktionsmodeller. Med hjälp av en sådan modell uppskattas förväntat värde för regressanden Y, givet värdena på regressorerna $X_1, \ldots, X_k$. Syftet är härvid att finna en modell som ger små prediktionsfel (skillnad mellan modellens predikterade värde och väntevärdet för givna värden på X-variablerna). En prediktionsmodell kan ofta betraktas som ett substitut som används i brist på bättre möjlighet d.v.s. en förklaringsmodell.

Vid modellbyggnaden bör man därför i princip ställa krav på modellen som om det vore fråga om en förklaringsmodell. Vid tolkning måste man däremot ta hänsyn till att det endast varit möjligt att konstruera en prediktionsmodell. Det finns flera tänkbara skäl till att det i en survey många gånger inte går att bygga en förklaringsmodell. Man kan ofta inte i modellen inkludera alla variabler som man vet (eller tror) påverkar Y-variabeln. Antalet presumtiva regressorer kan vara

för stort, vissa variabler kan vara för dåligt mätta, medan andra inte är mätta alls. I surveys är de presumtiva förklaringsvariablerna alltid korrelerade. Ju starkare tendensen till linjär samvariation mellan två eller flera X-variabler är,desto svårare blir det att avgöra vilka variabler som bör ingå i modeller samt vilka effekter regressorerna har. Detta problem benämnes multikollinearitetsproblemet.
De skilda syftena vid byggande av förklaringsmodeller och prediktionsmodeller (prognosmodeller) sätter självfallet sin prägel på modellarbetet och analysgången. Vissa procedurer och mått som kan vara aktuella och lämpliga vid konstruktion av en typ av modeller är inte lika aktuella och i vissa fall inte ens adekvata vid konstruktion av andra typer av modeller.

I kap. 6.4 berörs därför multipel korrelation och determination som mycket ofta används och framför allt tolkas på ett helt felaktigt sätt.

## 6.2 Modellarbetets allmänna frågor

Inför och under modellbyggnadsarbetet måste en rad frågor få sina svar.

a. Vilka variabler kan tänkas påverka responsvariabeln Y ?

b. Finns det en ömsesidig påverkan mellan Y och en eller flera andra variabler ? Om svaret på denna fråga är "ja" krävs i allmänhet en modell i form av ett simultant ekvationssystem i vilket varje ekvation kan sägas utgöra en delmodell. Hela modellen måste skattas simultant. Det är angeläget att kunna identifiera denna situation. Behandlingen av denna modelltyp ligger emellertid utanför ramen för denna framställning. (Se Hanushek and Jackson: Statistical Methods for Social Scientists [11] eller Kmenta: Elements of Econometrics [18].

c. Skall Y eller en funktion av Y utgöra responsfunktion ? Särskilt om Y är en kategorisk variabel (ofta 0-1 variabel som anger eventuell förekomst av viss egenskap) är speciella responsfunktioner aktuella. (Se kap. 7.)

d. Skall regressionsmodellen vara linjär i parametrarna och linjär i de presumtiva förklaringsvariablerna $X_1,\ldots,X_k$? Skall man införa produkttermer som representerar interaktion (samspel) och innebär att en förklaringsvariabels effekt varierar med nivån på en eller flera andra förklaringsvariabler ?

e. Skall regressionsfunktionen vara ickelinjär i parametrarna och i så fall av vilken form ?

f. Kan Y-observationerna antas ha samma varians kring sitt betingade väntevärde för olika värden $X_1,\ldots,X_k$ ? Om svaret är "nej" skall då detta föranleda övergång till generaliserad minstakvadratskattning ? (I de flesta fall torde antagandet om konstant varians i multipla regressionsmodeller leda till godtagbara analysresultat.)

## 6.3 Hjälpmedel vid modellbyggnad

Modellbyggnaden torde i de flesta fall starta med modeller som är linjära i både parametrar och variabler. I vissa fall börjar man med en modell som innehåller få regressorer i andra fall prövas först modellen med samtliga X-variabler som regressorer. Oavsett startpunkt har man vissa instrument till hjälp för att utvärdera olika modeller.

- plottar av residualmönster (se kap. 5.1)
- statistiska test
- enkla mått som karakteriserar modellen (determination, reststandardavvikelse, Mallows $C_p$)

Test. Konstruktion av statistiska test i samband med regressionsanalys återfinnes i läroböcker i regressionsanalys och ekonometri och skall inte beröras här. Användning och tolkning av vissa test missuppfattas emellertid ofta, varför det är befogat med en kort översikt.
I princip all estimation av modeller och alla statistiska test i regressions- och ekonometrilärоböcker förutsätter obundet slumpmässigt urval (OSU) av individer. Teori för andra urvalsformer lyser nästan helt med sin frånvaro.
Om urvalssannolikheten för individer varierar bör man vanligen väga med inverterade värdet av urvalssannolikheterna vid estimation och test. Vägningen förhindrar emellertid inte att variansen för estimatorer och testvariabler i de allra flesta situationer systematiskt underskattas för urval som ej är OSU. Följden blir att man erhåller för korta konfidensintervall i förhållande till nominella konfidensnivåer och att de reella felriskerna vid test blir högre än de nominella felriskerna. Annorlunda uttryckt löper man ökad risk för falska signifikanser.

I en utskrift från ett program för multipel linjär regressionsanalys förekommer flera olika test av modell och modellparametrar.

Olika typer av regressionsprogram. Det finns en typ av program i vilka man specificerar en modell som skattas. Modellen och dess parametrar utsätts sedan för olika test. Representanter för denna programtyp utgör SAS-programmet GLM och BMDP-programmet P1R. Det finns en annan typ av program i vilka man endast anger responsvariabel och en lista med potentiella regressorer, bland vilka vissa eventuellt är obligatoriska. Vid denna stegvisa regressionsanalys söker sig programmet fram till en "bästa" modell enligt ett av flera olika möjliga kriterier. Härvid startar man antingen med en minsta möjliga modell och lägger till en variabel i taget eller med modellen med alla potentiella regressorer och tar bort en variabel i taget. Det är i fallet då man lägger till variabler ofta också möjligt att ompröva tidigare beslut och ta bort en variabel innan man fortsätter med följande steg. Det har vuxit upp en rik flora av termer för det i grunden enhetliga partiella F-test som används i samband med modellbyggnaden enligt olika procedurer.

## 6.4 Test av hela modellen

Låt

$$Y = \alpha+\beta_1 X_1+\ldots+\beta_k X_k+\varepsilon \qquad (6.1)$$

med standardantaganden om $\varepsilon$, beteckna modellen som testas.

Det första test som påträffas i en programutskrift återfinnes i en liten varianstablå som kan se ut på olika sätt och som tyvärr också använder olika terminologi. I SAS-programmet GLM ser layouten ut på följande sätt. Exemplet är lånat från SAS User's Guide (flertalet värden är utelämnade här) dit läsaren kan vända sig för att erhålla en utförligare beskrivning.

Dependent variable: Oxy.

| Source | Df | Sum of squares | Mean square | F-value | Pr > F |
|---|---|---|---|---|---|
| Model | 6 | | | 18.67 | 0.0001 |
| Error | | | | | |
| Corrected total | | | | | |

Frihetsgradtalet (df) är 6 eftersom modellen innehåller 6 regressorer.
Nollhypotesen är $H_0: \beta_1=\beta_2=\ldots=\beta_6=0$ d.v.s. det finns inget linjärt regressionssamband mellan responsvariabeln och $X_1, X_2, \ldots, X_6$. Sannolikheten för ett så stort F-värde som det observerade eller större än enligt högerkolumnen 0.0001 om $H_0$ är sann. Således dras självfallet slutsatsen att en eller flera regressorer har en koefficient som är skild från 0. Om man vid ett sådant modelltest inte får signifikans enligt egen vald testnivå måste detta tolkas som att modellen inte är relevant. De följande testen i utskriften vilka avser enskilda parametrar saknar då mening. Däremot kan det naturligtvis finnas annan relevant information i utskriften, exempelvis residualplottar.
I motsvarande BMDP-program P1R skulle varianstablån använda en annan terminologi. "Model" är utbytt mot "Regression" och "Error" är utbytt mot "Residual".

Analysis of variance

| | Sum of squares | Df | Mean square | F-ratio | P(tail) |
|---|---|---|---|---|---|
| Regression | | 6 | | 18.67 | 0.0001 |
| Residual | | | | | |

## 6.5 Partiellt F-test och modelljämförelse

Det partiella F-testet är konstruerat för jämförelse av två regressionsmodeller. Den ena modellen har regressorerna $X_1, \ldots, X_p$

$$Y = \alpha+\beta_1X_1+\ldots+\beta_pX_p+\varepsilon \qquad (6.2)$$

Den andra större modellen innehåller därutöver ytterligare q regressorer $X_{p+1}, \ldots, X_{p+q}$

$$Y = \alpha+\beta_1X_1+\ldots+\beta_pX_p+\beta_{p+1}X_{p+1}+\ldots+\beta_{p+q}X_{p+q}+\varepsilon \qquad (6.3)$$

Man önskar testa hypotesen att den större modellen inte ökar förklaringsgraden signifikant eller med andra ord att $\beta_{p+1}=\ldots=\beta_{p+q}=0$. Om modellerna skiljer sig åt med avseende på en X-variabel (p=1) finns aktuella test i utskrifterna till vilket program som helst som skattar den större modellen. (Se nedan) Det finns ofta möjlighet att genomföra ett allmänt sådant test för $p \geq 2$ genom att i ett program specificera

en viss matris. Detta kan ofta vara svårt för den mindre rutinerade användaren. Det är emellertid möjligt att låta programmet skatta varje modell för sig och därefter beräkna F-kvoten manuellt enligt

$$F = \frac{[SS_{regr}(större) - SS_{regr}(mindre)]/g}{MS_{error}(större)} \qquad (6.4)$$

där SS betyder Sum of Squares och MS betyder Mean square. Frihetsgradtalen för F-testet är q respektive $df_{error}$ där det senare finns angivet i utskriften för den större modellen.
Ofta är det aktuellt att jämföra två modeller, den ena med och den andra utan en grupp regressorer bestående av indikatorvariabler för en viss kategoriindelning (som exempelvis socialgrupp). Kap. 13 i Kleinbaum and Kupper [17] ger ett flertal exempel på sådana test i samband med en modellbyggnadsprocess.

Exempel. I kap. 7 konstrueras modeller för att beskriva hur individers utövande av politisk aktivitet genom partier (A) varierar med bakgrundsvariabler: KÖN, ålder (indikatorvariablerna ÅLD2 och ÅLD3 ger uppdelning i tre åldersklasser) UTBildning över den obligatoriska, POLitiskt engagemang i föräldrahemmet. Alla variabler är binära och antar värden 0 eller 1. Det finns en rad speciella problem förknippade med modellens användning. Dessa diskuteras i kap. 7. Här är syftet enbart att illustrera testprocedurerna.
(Nominella signifikansnivåer stämmer inte med de reella)

Skrivsättet

A = KÖN ÅLD2 ÅLD3 UTB POL POL*KÖN UTB*POL UTB*ÅLD2 UTB*ÅLD3
POL*ÅLD2 POL*ÅLD3 (6.5)

som är vanligt för standardprogram anger modellen

$$A = \alpha + \beta_1 KÖN + \ldots + \beta_{11} POL \cdot ÅLD3 + \varepsilon \qquad (6.6)$$

Vid skattningen används som vikt inverterade urvalssannolikheter korrigerade för svarsbortfall (V544). SAS-programmet GLM ger varianstablån i tabell 6.1.

Tabell 6.1. Utdrag ur SAS GLM-program (avrundade tal).

Dependent variable: A

Weight: V544

| Source | Df | Sum of squares | Mean square | F-value | Pr>F | R-square |
|---|---|---|---|---|---|---|
| Model | 11 | 2791.41 | 253.76 | 18.06 | 0.0001 | 0.072 |
| Error | 2440 | 34291.01 | 14.05 | | | |
| Corrected total | 2451 | 37082.42 | | | | |

Modellen som helhet är som synes starkt signifikant vilket visas av att Pr>F är 0.0001.
För att undersöka om samspelet mellan ålder och politiskt engagemang är signifikant skattas även en modell som skiljer sig från ovanstående genom att termerna POL·ÅLD2 och POL·ÅLD3 saknas.

A = KÖN ÅLD2 ÅLD3 UTB POL POL·KÖN UTB·POL UTB·ÅLD2 UTB·ÅLD3

Varianstablån ges i tabell 6.2.

Tabell 6.2. Utdrag från SAS GLM-program (avrundade tal).

Dependent variable: A

Weight: V544

| Source | Df | Sum of squares | Mean square | F-value | Pr>F | R-square |
|---|---|---|---|---|---|---|
| Model | 9 | 2611.89 | 290.21 | 20.56 | 0.0001 | 0.070 |
| Error | 2442 | 34470.53 | 14.12 | | | |
| Corrected total | 2451 | 37082.42 | | | | |

Även denna modell är starkt signifikant med vanliga kriterier.

(Här är Pr>F angiven till 0.0001)

Hypotesen att de sanna regressionskoefficienterna för samspelstermerna POL·ÅLD2 och POL·ÅLD3 båda är 0 testas enligt formeln (6.4). Kvadratsumman för regression kallas i SAS-terminologi kvadratsumma för modell.

Eftersom q=2 regressorer slopats i den mindre modellen erhålles med hjälp av de båda varianstablåerna

$$F = \frac{(2791.41 - 2611.89)/2}{14.12} = 6.36$$

Enligt F-tabell för frihetsgradtalen 2 och 2442 (slå på 2 och $\infty$) svarar detta mot att Pr>F ligger mellan 0.001 och 0.005. Eftersom testförutsättningarna är mycket dåligt uppfyllda är den verkliga sannolikheten troligen större. Interaktionen torde dock kunna betraktas som reell vilket innebär att den mindre modellen förkastas.

## 6.6 Många namn för samma test i olika situationer

I programutskrifter förekommer flera olika test av enskilda koefficienter. Mest känt är förmodligen det partiella t-testet. (Se tabell 6.3)

Tabell 6.3. Utdrag ur SAS GLM-program (avrundade tal).

| Parameter | Estimate | T for H0: Parameter=0 | Pr>!T! |
|---|---|---|---|
| Intercept | 0.0607 | 2.39 | 0.0170 |
| X1 | 0.1506 | 3.32 | 0.0009 |
| X2 | 0.0891 | 3.14 | 0.0017 |
| X3 | 0.0561 | 1.84 | 0.0655 |
| X4 | 0.0677 | 2.24 | 0.0254 |
| X5 | -0.0423 | -2.50 | 0.0125 |
| X6 | -0.1163 | -3.30 | 0.0010 |
| X7 | -0.0663 | -1.70 | 0.0899 |
| X8 | 0.1418 | 3.49 | 0.0005 |
| X9 | 0.0471 | 0.97 | 0.3337 |
| X10 | -0.0111 | -0.30 | 0.7650 |
| X11 | 0.1010 | 2.49 | 0.0129 |

Ett intercept ingår alltid i en modell utom i speciella situationer, varför första radens test saknar intresse. På X1-raden jämföres en modell med alla elva X-variablerna som regressorer med en modell som inkluderar alla X-variablerna utom $X_1$ som regressorer. Nollhypotesen är således $H_0: \beta_1 = 0$ .

Med en konventionell testnivå på 0.05 skulle således $H_0$ för-

kastas eftersom sannolikheten att erhålla ett t-värde som till sitt absolutbelopp är större än det observerade 2.39 är 0.0170.

På X2-raden testas den fullständiga modellen mot en modell med 10 regressorer som utesluter endast X2 ($H_0:\beta_2=0$).

Av tabell 6.2 framgår att det finns flera modeller med 10 regressorer som ej är signifikant sämre än modellen med 11 regressorer. Dessa är modellerna som saknar enbart X3, som saknar enbart X7, som saknar enbart X9 respektive som saknar enbart X10. Observera att man inte, vilket ofta sker, kan dra slutsatsen att ingendera av X3, X7, X10 och X11 signifikant ökar förklaringsgraden och därför samtliga skulle kunna slopas i modellen.
Det kan mycket väl förhålla sig så att till exempel X3 ger signifikans om man jämför en modell som saknar både X3 och X10 med en modell som saknar enbart X10. (Om man överväger att på en gång utesluta X3, X7, X9 och X10 ur modellen måste man först estimera en modell utan dessa variabler och sedan genomföra det tidigare beskrivna testet enligt 6.4).

Partiella F-test. Ett partiellt F-test som är fullständigt exvivalent med det partiella t-testet återfinnes i GLM-utskrifterna omedelbart före t-testdelen (se tabell 6.4). Testet baseras på en kvadratsumma som i SAS-terminologi kallas Type IV SS. (Det observerade F-värdet är lika med kvadraten på t-värdet. För X1 gäller att $t^2=3.32^2=12.02$ medan F på grund av att beräkningarna är avrundade blir 12.03).

F-testet knutet till Type I SS (se tabell 6.4) brukar räknas till sekvensiella F-test beroende på att man genomför ett vanligt partiellt F-test för en sekvens av modeller med ökande (eller minskande) antal regressorer.
Vilken sekvens av modeller som studeras avgöres av MODEL-satsen i programmet. Specifikationerna

```
MODEL Y = X1  X2  X3;          (6.7)
```

och

```
MODEL Y = X3  X1  X2;          (6.8)
```

ger upphov till samma analys i GLM-programmet med undantag endast för beräkningen av Type I SS och tillhörande sekvens av partiella F-test. Dessa test knyts till följande

Tabell 6.4. Utdrag ur SAS GLM-program (avrundade tal)

| Source | Df | Type I SS | F-value | Pr>F | Df | Type IV SS | F-value | Pr>F |
|---|---|---|---|---|---|---|---|---|
| X1 | 1 | 968.82 | 68.94 | 0.0001 | 1 | 155.01 | 11.03 | 0.0009 |
| X2 | 1 | 311.16 | 22.14 | 0.0001 | 1 | 138.49 | 9.85 | 0.0017 |
| X3 | 1 | 83.86 | 5.97 | 0.0146 | 1 | 47.73 | 3.40 | 0.0655 |
| X4 | 1 | 584.95 | 41.62 | 0.0001 | 1 | 70.30 | 5.00 | 0.0254 |
| X5 | 1 | 319.89 | 22.76 | 0.0001 | 1 | 87.85 | 6.25 | 0.0125 |
| X6 | 1 | 164.54 | 11.71 | 0.0006 | 1 | 153.47 | 10.92 | 0.0010 |
| X7 | 1 | 15.65 | 3.96 | 0.0467 | 1 | 40.47 | 2.88 | 0.0899 |
| X8 | 1 | 155.04 | 10.89 | 0.0010 | 1 | 170.86 | 12.16 | 0.0005 |
| X9 | 1 | 16.35 | 1.16 | 0.2808 | 1 | 13.14 | 0.93 | 0.3337 |
| X10 | 1 | 46.14 | 3.28 | 0.0701 | 1 | 1.26 | 0.09 | 0.7650 |
| X11 | 1 | 87.01 | 6.19 | 0.0129 | 1 | 87.01 | 6.19 | 0.0129 |

sekvens av modeller för specifikation (6.7)

$$Y = \alpha+\varepsilon \qquad (6.9a)$$

$$Y = \alpha+\beta_1 X_1+\varepsilon \qquad (6.9b)$$

$$Y = \alpha+\beta_1 X_1+\beta_2 X_2+\varepsilon \qquad (6.9c)$$

$$Y = \alpha+\beta_1 X+\beta_2 X_2+\beta_3 X_3+\varepsilon \qquad (6.9d)$$

Ordningsföljden bestäms av ordningsföljden mellan regressorer i MODEL-satsen (6.7). F-testet på rad X1 i tabell 6.4 är därför ett test som jämför modell (6.9a) med modell (6.9b) enligt $H_0:\beta_1=0$. Ett tillräckligt högt F-värde innebär att modell (6.9b) är signifikant "bättre" än modell (6.9a).
F-testet av typ I på rad X2 jämför modellerna (6.9c) och (6.9b) med nollhypotesen $H_0:\beta_2=0$.

På detta sätt jämföres modeller med ökande antal termer. Observera att jämförelserna göres på samma sätt oavsett hur hög eller låg sannolikhet som erhålles i ett enskilt test. Om man testar med signifikansnivån 0.10 erhålles enligt tabell 4 en signifikant förbättring av modellen för var och en av variablerna $X_1,\ldots,X_8$ som läggs till. $X_9$ ger ingen signifikant förbättring. Trots detta jämföres på raden X10 modellen Y=X1...X9 med modellen Y=X1...X9 X10. Den senare modellen är signifikant "bättre" än den förra på signifikansnivån 0.10. Huruvida det blir en signifikant försämring eller ej om den större modellen reduceras till X1...X8 X10 genom att man tar bort X9 kan ej utläsas av tabell 4.
Det inträffar ofta att en variabel (egentligen koefficienten för variabeln) som är signifikant i en mindre modell blir insignifikant i en större modell. Exempelvis kan X7 eller X3 visa sig insignifikanta i modellen Y=X1...X8 trots att X3 varit signifikant i modellen Y=X1 X2 X3 och X7 varit signifikant i modellen Y=X1...X7.

F-test i stegvisa program. I de olika programmen för att stegvis bygga en regressionsmodell prövar programmet en sekvens av modeller. Vilken variabel som skall tillföras eller tas bort i det enskilda steget avgöres av något optimalitets-

kriterium. Det finns flera olika kriterier och flera olika sökprocedurer. Se emempelvis Draper and Smith [3 ] eller Younger [31] samt BMDP- och SAS-manualerna för programmet P2R respektive STEPWISE. I dessa program används partiella F-test som går under olika beteckningar. Efter varje steg i P2R har en modell med vissa regressorer skattats. Tabell 6.5 visar ett utdrag av vissa kolumner från en del av exemplet på sid 402 i BMDP-manualen 1979.

Tabell 6.5. Utdrag ur BMDP-manualen för programmet P2R (avrundade tal).

Step no. 3
Variable entered

| | |
|---|---|
| Multiple R | 0.4750 |
| Multiple R-square | 0.2256 |
| Adjusted R-square | 0.2124 |
| Standard error of est. | 37.9329 |

Analysis of variance

| | Sum of squares | Df | Mean square | F-ratio |
|---|---|---|---|---|
| Regression | 73789 | 3 | 24596.4 | 17.09 |
| Residual | 253246 | 176 | 1439.9 | |

| Variables in equation | | | Variables not in equation | |
|---|---|---|---|---|
| Variable | Coefficient | F to remove | Variable | F to enter |
| (Y-intercept -49.738) | | | | |
| AGE | 1.4359 | 23.99 | HEIGHT | 1.221 |
| CALCIUM | 20.7920 | 11.67 | WEIGHT | 0.004 |
| URICACID | 6.3444 | 5.84 | BRTHPILL | 1.655 |
| | | | ALBUMIN | 0.185 |

Det första F-testet i varianstablån är det vanliga testet av hela modellen Y=AGE CALCIUM URICACID.
F-värdet 17.09 kan till exempel jämföras med kritiska värdet 3.78 för test på signifikansnivån 0.01. Modellen är således signifikant.
"F to remove" är samma test som typ IV testet i GLM (tabell 6.4 ovan).
F-värdet 23.99 ger således ett test av $H_0: \beta_{AGE}=0$ i modellen

ovan. Frihetsgradtalen är 1 och 176. Kritiskt värde på nivån 0.01 är 6.63 varför $H_0$ förkastas. F-värdet 11.67 ger ett test av $H_0:\beta_{CALCIUM}=0$ för samma trevariabelmodell som ovan.

Testen med hjälp av "F to enter" avser olika fyravariabelmodeller nämligen de som erhålles genom att lägga till en variabel i trevariabelmodellen ovan. F-värden 1.221 på raden HEIGHT ger ett test av $H_0:\beta_{HEIGHT}=0$ avseende modellen

Y = AGE CALCIUM URICACID HEIGHT.

F-värdet 0.0004 på raden WEIGHT ger på samma sätt ett test av $H_0:\beta_{WEIGHT}=0$ avseende modellen

Y = AGE CALCIUM URICACID WEIGHT.

Frihetsgradtalen är för båda testen 1 och 175. Regressionskoefficientskattningarna är insignifikanta både på 1 %-nivån och 5 %-nivån (kritiska värden 6.63 resp. 3.84). Vid bestämning av signifikansgränser för test vid stegvis regression bör man komma ihåg att den verkliga risken att felaktigt förkasta $H_0$ är mycket större än den nominella risk som svarar mot F-tabellens värden. Detta beror på att programmet i varje steg väljer en variabel av flera möjliga för att införa i modellen (eller ta bort från modellen). Se kap. 6.7 samt BMDP manualen 1979 sid 403 och 855 för vidare diskussion.

## 6.7 Signifikant modell visavi signifikant prediktionsförmåga

Det har tidigare påpekats (kap.1.3) att sökandet efter en modell vars form påverkas av datamaterialet, leder till överanpassning. De verkliga signifikansnivåerna är inte så höga (riskerna att felaktigt förkasta $H_0$ är inte så låga) som de nominella signifikansnivåerna. Detta är kanske mest uppenbart vid användning av program för stegvis regressionsanalys. I varje steg väljer programmet en variabel som införes i (eller borttas från) modellen bland ett ofta stort antal kandidater. Signifikansgränserna bör därför höjas långt över de i F-tabellen angivna värdena. (Se BMDP-manualen 1979 sid. 403 och 855).
En annan fråga är i vilken utsträckning ett statistiskt signifikant resultat även är praktiskt signifikant, d.v.s. av praktiskt värde. Ett test förmåga att upptäcka en viss av-

vikelse från nollhypotesen (testets styrka) ökar med ökande urvalsstorlek. Med stora datamaterial kommer därför även små avvikelser att bli statistiskt signifikanta. Variationsvidden för predikterade Y-värden måste vara av betydelse i förhållande till slumpvariationen av Y-värden för fixa värden på regressorerna. Draper and Smith [3 ] (sid. 93 och sid. 129) anser att det observerade F-värdet bör överstiga minst 4 gånger det F-värde som enligt F-tabellen svarar mot den valda signifikansnivån.

## 6.8 Determinationsmåttets tolkning och användning

Multipel korrelation och determination är meningsfulla mått endast i surveys, inte i experimentella undersökningar (se kap. 2.2).
Vid modellbyggnad inriktas ansträngningarna ofta alltför ensidigt mot att finna en modell med så hög determination som möjligt.
I många undersökningar tas en hög multipel korrelation R eller en hög determination $R^2$ (som kan anta värden i intervallet [0,1]) till intäkt på att modellen är bra, oavsett vad den skall användas till. En hög determination innebär emellertid endast att observationernas variation kring de enligt modellen predikterade värdena är betydligt mindre än totala variationen Y.
Det följer inte som en konsekvens härav att man får en riktig bild av vilka variabler som styr Y-variabeln och att dessas effekter är adekvat uppskattade.
Urvalets totala kvadratavvikelsesumma för Y kring sitt totalmedelvärde $\bar{Y}$ kan delas upp i två komponenter.
Den ena mäter hur mycket de predikterade värdena $\hat{Y}$ enligt den skattade regressionsfunktionen varierar kring totalmedelvärdet $\bar{Y}$ ("förklarad" variation). Den andra mäter Y-värdenas variation kring predikterade $\hat{Y}$-värden (restvariation eller "oförklarad" variation). Storleken av dessa båda komponenter jämfört med totala variationen kan uttryckas med hjälp av urvalets multipla korrelationskoefficient R.
Den "förklarade" variationen (determinationen) uppgår till andelen $R^2$ och den "oförklarade" variationen (restvariationen) uppgår till andelen $1-R^2$ av totala variationen.

$$\Sigma(y_i-\bar{y})^2 = \Sigma(y_i-\hat{y}_i)^2+\Sigma(\hat{y}_i-\bar{y})^2 =$$

$$= (1-\hat{R})^2\Sigma(y_i-\bar{y})^2+\hat{R}^2\Sigma(y_i-\bar{y})^2 \qquad (6.10)$$

Determination vilseledande. Frågan är om det finns andra mått inom statistiken som är så missuppfattade och missbrukade som olika korrelationskoefficienter, inte minst den multipla korrelationskoefficienten.
Determinationskoefficienten $R^2$ har olyckligtvis fått benämningen "förklaringsgrad". Man kan säga olyckligtvis på grund av att hög determinationsgrad inte säger ett dyft om huruvida modellen ens är rimlig. En helt korrekt förklaringsmodell kan å andra sidan ha hur låg determination som helst. Det kan ju mycket väl vara så att större delen av variationerna i Y är slumpmässiga. $R^2$ är helt enkelt ett deskriptivt mått som kan karakterisera storleksförhållandena mellan restvariation och total variation för förklaringsmodeller såväl som för prediktionsmodeller och kontrollmodeller.
Determinationen bestäms inte enbart av hur stark effekt (hur höga regressionskoefficienter) olika förklaringsvariabler har utan också av hur vanliga olika kombinationer av värden på förklaringsvariablerna är. Ett enkelt exempel belyser detta.

Exempel 6.1. Inom två branscher A och B har vardera 800 personer valts ut i ett obundet slumpmässigt urval. Bransch A har jämn fördelning av män och kvinnor medan bransch B är mycket starkt mansdominerad. Inom de båda branscherna finner man exakt samma förhållande vad beträffar partisympatier. En fjärdedel av kvinnorna och hälften av männen sympatiserer med socialistiska blocket.

Bransch A

| Sympatier | Kön: Kv (x=0) | Kön: Män (x=1) |
|---|---|---|
| Soc. (Y=1) | 100 | 200 |
| Borg. (Y=0) | 300 | 200 |

Bransch B

| Sympatier | Kön: Kv (x=0) | Kön: Män (x=1) |
|---|---|---|
| Soc. (Y=1) | 25 | 350 |
| Borg. (Y=0) | 75 | 350 |

Data enligt tabellen ger samma regressionslinje, $y = 0.25+0.25x$ för regressionen av partisympati på kön, för branscherna A och B. Determinationen blir emellertid $r^2=0.063$ för bransch A men endast $r^2=0.027$ för bransch B. Denna paradox förklaras helt av att könsfördelningen skiljer sig åt mellan branscherna.

Vi ser också att för kvalitativa variabler blir determinationen ytterst låg trots att förekomsten av den ena egenskapen (socialistisk partisympati) varierar starkt med förklaringsvariabeln (kön).

De flesta användare av determinationsmåttet har förmodligen inte ett ögonblick reflekterat över att kvadratavvikelser (och varianser) inte är särskilt naturliga spridningsmått och inte heller över att det finns andra mått på "förklaringsgrad" som ger en helt annan, inte lika positiv bild. Anledningen till att varianser (standardiserade kvadratavvikelsesummor) dominerar som spridningsmått inom statistiken är inte att de är intuitivt tilltalande, utan att de är matematiskt hanterliga vid härledning av statistisk teori. Vid tillämpning av statistisk metodik på empiriskt material är en enkel funktion av variansen, nämligen standardavvikelsen av mycket större intresse.

Förklarad standardavvikelse. Varför utgår man då inte från standardavvikelsen vid konstruktion av ett mått på en regressionsmodells förklaringsgrad ? Förklaringen är att standardavvikelsen inte additivt kan delas upp i enkla komponenter analogt med uppdelningen av kvadratavvikelsesumman ovan. Däremot kan man naturligtvis för viss given multipel korrelation jämföra reststandardavvikelsen i Y med totala standardavvikelsen i Y. Approximativt gäller för stickprovsvarianserna

$$s_{y\cdot x}^2 = (1-R^2)s_y^2 \qquad (6.11)$$

vilket ger motsvarande samband för standardavvikelserna

$$s_{y\cdot x} = s_y\sqrt{1-R^2} \qquad (6.12)$$

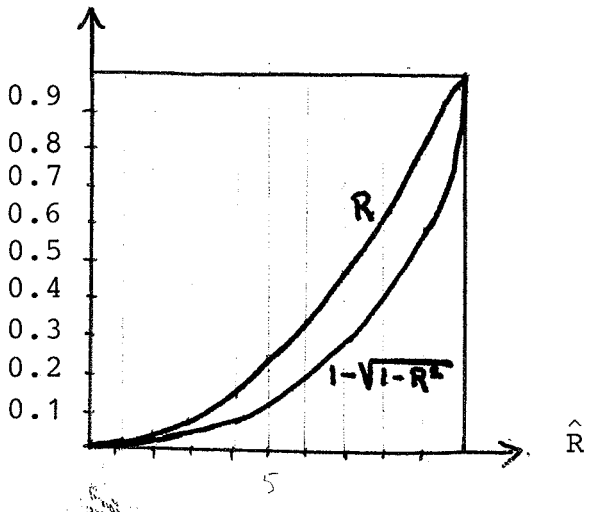
Man skulle kunna säga att av ursprunglig variation i Y kan andelen $\sqrt{1-R^2}$ inte "förklaras". Däremot har modellen "förklarat" andelen $1-\sqrt{1-R^2}$ av ursprunglig variation i Y. Det

införda måttet "förklarad standardavvikelse" är inte etablerat inom statistiken, men det belyser svagheten i måttet "förklarad varians" som ofta (för att inte säga vanligen) blir direkt vilseledande för statistikanvändare. Se tabell 6.6 och figur 6.1.

Tabell 6.6 Förklaringsgrad hos regressionsmodell för olika multipla korrelationer.

| Multipel korrelation $R$ | Andel "förklarad varians" (determination) $R^2$ | Andel "förklarad standardavvikelse" $1-\sqrt{1-R^2}$ |
|---|---|---|
| 0.10 | 0.01 | 0.005 |
| 0.20 | 0.04 | 0.020 |
| 0.30 | 0.09 | 0.046 |
| 0.40 | 0.16 | 0.083 |
| 0.50 | 0.25 | 0.134 |
| 0.60 | 0.36 | 0.200 |
| 0.70 | 0.49 | 0.286 |
| 0.80 | 0.64 | 0.400 |
| 0.90 | 0.81 | 0.564 |
| 0.95 | 0.903 | 0.689 |
| 0.98 | 0.960 | 0.800 |
| 0.99 | 0.980 | 0.859 |
| 0.995 | 0.990 | 0.900 |
| 0.999 | 0.998 | 0.955 |

Figur 6.1 Förklaringsgrad i termer av varians och Standardavvikelse.



Som synes blir förklaringsgraden mycket lägre i termer av standardavvikelse än i termer av varians. Exempelvis ger en multipel korrelation på 0.80 för varians en "förklaringsgrad" på 64 % men för standardavvikelse en "förklaringsgrad" på endast 40 %. För en multipel korrelation på 0.995 är motsvarande förklaringsgrader 99 % respektive 90 % (ej förklarad variation 1 % respektive 10 % !).

## 6.9 Stegvis ökning av determinationen

Vid modellarbete är det vanligt att man studerar följder av regressionsmodeller med ökande antal regressorer. För varje modell bestäms determinationen $R^2$. För en multipel regressionsmodell som är linjär i både parametrar och variabler kan varje variabels tillskott till determinationen korrekt mätas under en enda förutsättning, nämligen att regressorerna är okorrelerade. I urvals- eller populationsstudier föreligger aldrig denna förutsättning utan regressorerna är korrelerade.
Ofta påträffas tabeller av typ tabell 6.7 i redovisningar av modellers förklaringsvärde.

Tabell 6.7. Förklaringsvärde för olika modeller.

| Beroende variabel | Förklarings-variabler | Förklaringsgrad $\hat{R}^2$ | Ökning av förklaringsgraden |
|---|---|---|---|
| Y | $X_1$ | 0.8339 | 0.8339 |
| Y | $X_1, X_6$ | 0.8568 | 0.0229 |
| Y | $X_1, X_6, X_2$ | 0.8628 | 0.0060 |

Grunddata är hämtade ur Mendenhall and Reinmuth [21] sid. 492-493. $Y, X_1, X_2$ och $X_6$ anger försäljningspris på villor, bostadsyta, antal sovrum respektive förekomst av garage. Sådana tabeller åtföljs i många fall av påståenden såsom: $X_1$ förklarar 83.39 %, $X_6$ förklarar 2.29 % och $X_2$ förklarar 0.6 % av variationerna i Y. Påståendet är grundfalskt. Man får en helt annan bild av betydelsen av variablerna $X_1$ och $X_6$ om man i stället jämför tvåregressorsmodellen ovan med en modell med $X_6$ som enda regressor. (Tabell 6.8).

Tabell 6.8. Förklaringsgrad för olika modeller.
Samma grunddata som i föregående tabell.

| Beroende variabel | Förklarings-variabler | Förklaringsgrad (Determination) | Ökning av förklaringsgraden |
|---|---|---|---|
| Y | $X_6$ | 0.2574 | 0.2574 |
| Y | $X_1, X_6$ | 0.8568 | 0.5994 |

Som synes kan man inte dra några som helst slutsatser om de olika regressorernas relativa betydelse i dess påverkan av Y på basis av determinationsgraden för en följd av regressionsmodeller. Tillskottet till förklaringsgraden för $X_6$ och $X_1$ blir nu 25.74 % respektive 59.94 % mot 83.39 % och 2.29 % enligt föregående tabell.
Man får således helt skilda resultat beroende på i vilken ordning regressorerna införes i modellen. Överhuvud taget är det inte meningsfullt att studera ökningen av determinationen annat än för modeller som skall användas för prognoser. I detta fall användes inte ökningen av determinationen för att beskriva olika regressorers förklaringsvärde utan enbart för att beskriva hur determinationen ökar med antalet regressorer. Det är en allmän tendens för korrelerade regressorer att marginaleffekten av ökat antal regressorer avtar ytterst snabbt (vilket illustreras av tabell 6.7).

## 6.10 Determination vid förekomst av upprepade observationer

I vissa situationer förekommer upprepade observationer på Y för en given kombination av värden på X-variablerna. Y-observationernas variation kring den anpassade regressionsfunktionen $\hat{Y}$ (för en given kombination av värden på X-variablerna) kan då delas upp i gruppmedelvärdets $\bar{Y}_x$ avvikelse från $\hat{Y}$ och de enskilda observationernas avvikelse från gruppmedelvärdet $\bar{Y}_x$.
Även om alla gruppmedelvärden för Y skulle stämma exakt överens med den anpassade modellen blir inte determinationen $R^2=1$, såvida det inte helt saknas slumpvariation, d.v.s. alla upprepningar för givna värden på X-variablerna leder till identiska resultat.

Situationen med upprepade observationer på Y föreligger bl.a. då alla X-variablerna är kategoriska. Om dessutom Y är en binär variabel som endast kan anta värdena 0 och 1 (indikerande avsaknad eller förekomst av viss egenskap) anger väntevärdet för Y sannolikheten P(Y=1). För denna typ av linjära sannolikhetsmodeller kan man vänta sig att $R^2$-värdet ligger mycket lågt, i allmänhet under 0.2 och mycket ofta även under 0.1, även om alla predikterade sannolikheter skulle råka stämma exakt med den anpassade modellen. En tillämpning av detta slag (politisk aktivitet genom partier) behandlas i kapitel 8. En enkel illustration gavs i exempel 6.1.

## 6.11 Mallow's $C_p$

Mallow's $C_p$ är ett mått på totalt prediktionsfel för Y summerat över alla observationer. Det tar således hänsyn till både systematiskt fel förorsakat av en felspecificerad modell och till det slumpmässiga felet för prediktorn. För en korrekt specificerad modell har $C_p$ väntevärdet p där p anger antal parametrar (regressionskoefficienterna för regressorerna plus interceptet). För felaktigt specificerade modeller är väntevärdet större än p. Med alldeles för få regressorer i modellen ligger $C_p$ oftast långt över p. Ju närmre den sanna modellen den prövade modellen ligger desto närmre p bör i princip $C_p$ hamna. Kruxet är naturligtvis att man inte känner p. Måttet $C_p$ brukar därför plottas mot antalet parametrar i den prövade modellen för ett antal modellvarianter som övervägts. För en bra modell bör $C_p$ och p vara i stort sett lika stora. Se exempelvis Draper and Smith [3 ] kap.6 för närmare beskrivning och illustration av användningen. Vissa datorprogram ger möjlighet att beräkna $C_p$. I BMD-programmet för stegvis regression kan $C_p$ användas som kriterium för modellval.

## 6.12 Utelämnande av ointressanta variabler

I experimentella undersökningar göres vanligen förklaringsvariablerna ortogonala (varieras oberoende av varandra) med hjälp av en balanserad experimentplan (se kap. 1.4). Detta medför ur skattningssynpunkt två fördelar.

Antag att man har en sann regressionsmodell $Y = \alpha+\beta_1 X_1+\ldots+$ $+\beta_1 X_k+\varepsilon$. Den första fördelen består i att man då kan skatta regressionskoefficienterna väntevärdesriktigt för en delmängd av X-variabler, säg $X_1,\ldots,X_{k-p}$ där $1<p<k$, om man ansätter en modell som endast innehåller dessa X-variabler som regressorer. Detta är inte möjligt i surveys på grund av att alla förklaringsvariabler är korrelerade. Det förekommer ibland i surveys att man utelämnar vissa förklaringsvariabler från en regressionsmodell på grund av att man inte primärt är intresserad av dessa variablers effekt (vilken kanske är känd sedan tidigare) utan vill koncentrera sig på att undersöka andra variablers effekt. Detta är helt felaktigt och ger upphov till systematiska fel vid skattningen av regressionskoefficienter. Felets storlek beror av hur stark samvariationen är mellan den enskilda regressorn och utelämnade relevanta förklaringsvariabler. (Se kap. 5.3.) En grav variant av denna felaktiga analys består i att man beräknar regressionen av Y på en X-variabel i taget.
Den andra fördelen med balanserade experimentplaner består i att skattningarna av parametrar för olika regressorer blir oberoende. Detta är inte fallet i surveys vilket ger upphov till stora problem vilka behandlas i följande moment.

## 6.13 Multikollinearitetsproblemet

Uppsättningen potentiella regressorer är i surveys alltid korrelerade. Om någon X-variabel har hög multipel korrelation med övriga X-variabler föreligger det ett så kallat multikollinearitetsproblem.
Multikollinearitet i egentlig mening betyder funktionellt linjärt samband mellan en av X-variablerna och en eller flera andra X-variabler. Ordet har dock även kommit att användas för att beteckna situationen då det föreligger hög multipel korrelation mellan X-variabler. Här används det i denna bemärkelse.
Förekomsten av multikollinearitet innebär med Huangs [13] sammanfattning att

a. regressionskoefficienterna skattas mycket osäkert på grund av stor slumpmässig variation
b. specifikationen av modellen blir osäker med avseende på vilka variabler som skall inkluderas. Utelämnade relevanta variabler medför systematiska

skattningsfel (vilket diskuterats i föregående moment)

c. som konsekvens av punkterna 1 och 2 uppstår svårigheter vid tolkningen av de skattade regressionskoefficienterna.

Upptäckt av multikollinearitet. Modellbyggnaden bör alltid starta med att korrelationsmatrisen för Y och alla potentiella regressorer $X_1,\ldots,X_m$ beräknas. Om det finns ett eller flera par av X-variabler med hög inbördes enkel korrelation är detta ett direkt uttryck för att multikollinaritetsproblemet är allvarligt.

Det är självfallet svårt att ange gränser för hur höga enkla korrelationer som kan tolereras utan att skattningsproceduren helt spårar ur. Ju flera presumptiva regressorer som föreligger desto allvarligare effekt leder även ganska låga enkla korrelationer till. Tyvärr utgör frånvaron av höga enkla korrelationer ingen som helst garanti för att det saknas hög multipel korrelation mellan någon X-variabel och övriga X-variabler.

Bästa sättet att avgöra vilken effekt korrelationen mellan regressorerna har på variansen för skattningarna av regressionskoefficienterna är att studera inversen till korrelationsmatrisen för X-variablerna (observera att Y-variabeln ej skall ingå i matrisen).

Diagonalelementen, de så kallade "variance inflationary factors" (VIF), anger hur många gånger variansen för respektive parameterskattning ökas på grund av förekomsten av korrelation mellan regressorerna jämfört med om dessa vore okorrelerade. Marquart [20] säger i en artikel som diskuterar skattningsproblem vid förekomst av korrelerade X-variabler att den största VIF som kan tolereras bör vara större än 1.0 men med säkerhet inte så stor som 10.

En modell $E(Y) = \alpha+\beta_1X_1+\beta_2X_2$ ger följande variansinflation (som blir lika stor för $\beta_1$ och $\beta_2$) för olika värden på korrelationen $r_{12}$ mellan $X_1$ och $X_2$. (Tabell 6.9.)

Tabell 6.9. Exempel på variansinflation för modell med två regressorer.

| $r_{12}$ | VIF |
|---|---|
| 0.70 | 1.96 |
| 0.80 | 2.78 |
| 0.90 | 5.26 |
| 0.95 | 10.26 |
| 0.98 | 20.25 |

Det bör ånyo betonas att förekomsten av så höga enkla korrelationer får mycket allvarligare effekter i flerregressionsmodeller. En illustration av detta förhållande återfinnes i kap. 8.
Inversen till variansmatrisen kan i exempelvis SAS beräknas med hjälp av procedurerna CORR och MATRIX. (Se Exempel 6.2.)

Exempel 6.2. Beräkning av inversen till korrelationsmatrisen med SAS.
Beräkningen förutsättes avse den senast bildade datamängden i SAS-programmet.

```
PROC CORR NOPRINT NOSIMPLE OUTP=C;
   VAR X1 X2 X3 X4;
   WEIGHT V544;
PROC MATRIX;
   FETCH C;
   V = INV(C);
   PRINT V;
   TITLE BERÄKN AV INVERS KORRMATRIS
         MED VIKT = V544;
```

Som utskrift erhålles först korrelationsmatrisen och därpå den inverterade matrisen.

Tolerans av $R^2$ i datorprogram. BMDP-programmet P1R för linjär multipel regression har en spärr mot för hög multipel korrelation mellan regressorer. För att förhindra att det uppstår en determination $R^2$ mellan regressorer som är högre än 0.99 utesluts om nödvändigt en eller flera av de X-variabler som angivits i modelldeklarationen. Användaren av programmet har frihet att sätta denna toleransgräns lägre än 0.99 om så önskas. Motsvarande SAS-program GLM har ingen motsvarande sådan spärr.
BMDP-programmet P2R för stegvis regression har samma slags toleransgräns som P1R. Dessutom anges för varje regressor som inkluderats i modellen hur högt värdet är på toleransen definierad som TOLERANCE = $1-R^2$, där R anger regressorns multipla korrelation med övriga regressorer i modellen.

"Lösning" av multikollinearitetsproblemet. Då hög grad av multikollinearitet påträffas för en grupp X-variabler som logiskt sett borde ingå i regressionsmodellen står man inför ett svårt val. Om alla de aktuella X-variablerna med-

tages erhålles ytterligt osäkra skattningar av regressionskoefficienter på grund av stor slumpmässig variation. Om man å andra sidan utesluter en eller flera X-variabler som orsakar den höga multikollineariteten minskar variansen för koefficientskattningarna men å andra sidan uppstår risk för att det systematiska felet (bias) blir stort.
Den vanligaste åtgärden i denna situation torde vara att slopa den eller de variabler i modellen som orsakar multikollineariteten. Åtgärden ligger närmast till hands då syftet är prediktion av förväntade Y-värden.
Om storleken på regressionskoefficienterna är av primärt intresse vilket är fallet i förklarande undersökningar finns inte mycket som kan göras åt problemet. Man nödgas konstatera att hur man än beter sig är det inte möjligt att erhålla pålitliga koefficientskattningar.
Även om medelfelet för koefficientskattningarna är måttligt (de slumpmässiga variationerna måttliga) kan det systematiska felet (bias) vara stort på grund av att en eller flera utelämnade relevanta variabler har hög multikollinearitet med regressorerna.
En indikation (förutom stort medelfel) på att en viss regressors koefficient är mycket osäkert skattad är att koefficientens värde fluktuerar kraftigt då vissa av de tveksamma variablerna tillföres eller tas bort från modellen. Man hoppas således vid stegvis ökning av antalet regressorer i modellen att koefficientskattningarna för tidigare medtagna regressorer skall vara stabila från steg till steg, vilket utgör en indikation på (men inget bevis för) att respektive regressorers effekter är någorlunda riktigt skattade.

Koefficienttest. Test av enskilda regressionskoefficienter vid förekomst av hög multikollinearitet kan vara direkt vilseledande på grund av att koefficientskattningarna då blir mycket kraftigt korrelerade. Det är ofta nödvändigt att testa koefficienter i grupp med hjälp av det allmänna partiella F-testet.

## 6.14 Program för stegvis regression

Programmen för stegvis regression konstruerar modellen enligt ett av flera möjliga kriterier vilket vanligen kan väljas av

användaren. Låt säga att kriteriet är maximal förbättring av determinationen $R^2$ för varje steg.
Estimationsproceduren är utsatt för risken att en helt felaktig modell byggs upp om det förekommer multikollinearitet bland de presumptiva regressorerna.
En variabel som saknar effekt men har hög multipel korrelation med verkliga förklaringsvariabler kan ofta införas i modellen på ett tidigt stadium som "skuggvariabel" för de andra.
Antag att två verkliga förklaringsvariabler $X_1$ och $X_2$ är starkt positivt korrelerade och har kraftiga och ungefär lika stora effekter men med omvända tecken. Troligen skulle då regressionsproceduren inkludera den ena variabeln i modellen men inte den andra beroende på att den senare variabeln som infördes skulle höja determinationen obetydligt. Dessutom skulle utelämnandet av en variabel orsaka ett kraftigt systematiskt fel (bias) vid skattningen av effekten för den inkluderade variabeln. Hanushek and Jackson [11] är mycket hårda i sin kritik av stegvisa regressionsprocedurer:

> "Stepwise regression appears to promise something that it cannot deliver. It is not possible to use stepwise regression to give both the model and the parameter estimates. Nor is it possible to use either the order of entry into a stepwise procedure or the parameter estimates of intermediate stages to make inferences about the importance of particular variables (except in the context of one specific sample). To the extent that the purpose of estimation is to make inferences about population relationships on the basis of sample information, a stepwise procedure can be very misleading".

## 6.15 Avvikande smågrupper

En liten grupp individer med avvikande värden på en regressor skiljer sig ibland klart från i övrigt jämförbara individer vad gäller Y-värden. Det är då i sig önskvärt att beskriva detta i prediktionsmodellen. Uppskattningen av den aktuella regressorns parameter blir emellertid osäker och man kan dessvärre också orsaka större fel i uppskattningen av andra regressorers parametrar än om variabeln utelämnas ur modellen. (Ett exempel på detta dilemma ges i kap. 8 där socialgrupp I, som är liten, cirka 7 % av individerna i urvalet, visar annat mönster beträffande politisk aktivitet genom partier än övriga socialgrupper.)

## 6.16 Standardiserade regressionskoefficienter och relativa effekter

Man försöker ibland jämföra de olika regressorernas relativa betydelse för variationerna i Y-variabeln med hjälp av standardiserade regressionskoefficienter (ofta kallade β-koefficienter). Dessa är de vanliga regressionskoefficienterna beräknade på standardiserade variabler. Standardiserade regressionskoefficienten för en viss Y-variabel anger hur många standardavvikelser Y förväntas ändras då X-variabeln ökas med en standardavvikelse medan övriga variabler hålles konstanta.

Dessa koefficienter är emellertid inte så entydiga som namnet antyder. Storleken beror nämligen på variationerna för $X_j$ och Y i det speciella urvalet. Om det aktuella samplet uppvisar osedvanligt stor variation för $X_j$ blir också standardavvikelsen för $X_j$ och därmed den standardiserade gegressionskoefficienten ovanligt stor. Detta gäller alltså oavsett värdet på den skattade icke standardiserade regressionskoefficienten. Denna egenskap gör det meningslöst att jämföra standardiserade regressionskoefficienter från olika undersökningar, vilket däremot ej gäller de icke standardiserade koefficienterna.

Det är inte heller meningsfullt att använda standardiserade regressionskoefficienter om vissa X-variabler är kategoriska. Sådana kategoriska variabler anger vanligen förekomst eller avsaknad av en viss egenskap och kodas vanligen 1 eller 0. Regressionskoefficienten anger då skillnaden i förväntat värde för Y mellan individer som har och individer som saknar egenskapen. Eftersom endast värdena 0 och 1 (saknar/besitter egenskapen) är möjliga leder omräkning av effekten med hjälp av standardavvikelsen till en meningslöshet.

De icke standardiserade regressionskoefficienterna är vanligen att föredra. Detta gäller särskilt jämförelser av olika dikotoma variablers effekter och jämförelser (oavsett variabeltyp) av en viss variabels effekt i olika samples.

# 7. REGRESSIONSMODELLER FÖR BINÄR BEROENDE VARIABEL

## 7.1 Den linjära modellens svagheter

Låt $Y_i$ vara en binär beroende variabel som anger avsaknad av en viss egenskap ($Y_i=0$) eller förekomst av egenskapen ifråga ($Y_i=1$). Denna variabel kan ange val mellan två alternativ, såsom kollektivt eller privat transportmedel till arbetsplatsen, borgerligt eller socialistiskt parti vid riksdagsval. En annan vanlig situation är den där $Y_i$ indikerar om en individ företagit en viss aktivitet eller ej: ett kommunalt bibliotek har besökts under senaste året, en viss politisk aktivitet har förekommit t.ex. deltagande i politiskt möte.

En linjär regressionsmodell med en förklaringsvariabel skrivs

$$Y_i = \alpha+\beta X_i+\varepsilon_i \qquad (7.1)$$

där $X_i$= värde på förklaringsvariabeln (säg inkomst) för i:te individen

$$Y_i = \begin{cases} 1 & \text{om viss respons förekommer (säg privat transportmedel väljs)} \\ 0 & \text{om responsen saknas (kollektivt transportmedel väljs)} \end{cases}$$

$\varepsilon_i$ = oberoende stokastiska slumptermer med väntevärde 0 (i=1,...n)

Låt $P_i = P(Y_i=1)$, d.v.s. sannolikheten att $Y_i$ antar värdet 1.

Eftersom $E(\varepsilon_i=0)$ har vi

$$E(Y_i) = \alpha+\beta X_i \qquad (7.2)$$

men även

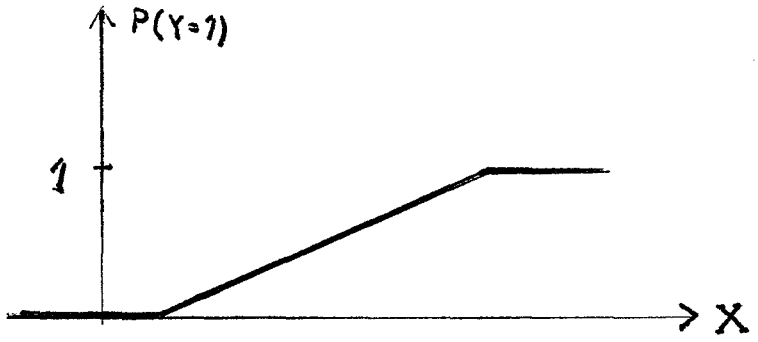
$$E(Y_i) = 1\cdot P(Y_i=1)+0\cdot P(Y_i=0) = P_i$$

Väntevärdet för regressionsmodellen anger således hur sannolikheten $P(Y_i=1)$ beror på förklaringsvariabeln $X_i$. Eftersom en sannolikhet är begränsad till intervallet [0;1] måste vi definiera $P_i$ på följande sätt:

$$P_i = \begin{cases} 0 & \text{då } \alpha+\beta X_i<0 \\ \alpha+\beta X_i & \text{då } 0\leq\alpha+\beta X_i\leq 1 \\ 1 & \text{då } \alpha+\beta X_i> \end{cases} \qquad (7.3)$$

Grafiskt beskrivs sannolikhetsmodellen av figur 7.1 (som förutsätter att $\beta>0$ d.v.s. att P(Y=1) ökar med X).

Figur 7.1. Linjär sannolikhetsmodell.




Variansen för en variabel $Y_i$ som antar värdena 0 och 1 med sannolikheter $P_i$ respektive $1-P_i$ är $P_i(1-P_i)$. Eftersom $Y_i$ har en systematisk (fix) del $\alpha+\beta X_i$ och en stokastisk del $\varepsilon_i$ enligt modellen (7.1) blir variansen för feltermen (givet $X_i$)

$$\text{Var}(\varepsilon_i) = \text{Var}(Y_i) = P_i(1-P_i) \qquad (7.4)$$

där $P_i = \alpha+\beta X_i$ . Variansen för $\varepsilon_i$ är följaktligen inte konstant, vilket är fallet för standardmodellen för linjär regression, utan varierar med $X_i$ .

Om modellen skattas med vanlig ovägd minstakvadratmetod (OLS) kan vi råka ut för att $\hat{Y}_i$ för vissa $X_i$ blir <0 och för andra $X_i$ blir >1. Detta kan i och för sig avhjälpas genom att sätta $\hat{Y}_i$ till 0 respektive 1 för dessa fall. Ofta vet vi emellertid att $Y_i=1$ inte är en omöjlig respektive säker händelse för dessa X-värden. I många fall förekommer observationer i samplet som bekräftar detta.

För många fördelningar av X i urvalet (en förutsättning är att man erhållit åtminstone något X-värde <a och/eller >b är minstakvadratestimatorn av $\beta$ biased. Oavsett fördelningen av X-värden är estimatorn inte effektiv i klassen av linjära estimatorer på grund av att variansen för $\varepsilon_i$ ej är konstant. Man skulle därför kunna förvänta sig att vägd minstakvadratmetod vore att föredra framför ovägd. Den förra kan emellertid inte tillämpas direkt eftersom $P_i$ är okänt. I ett första steg måste därför modellen skattas med den ovägda minstakvadratmetoden i syfte att erhålla skattningar

$$\hat{\sigma}^2_{\varepsilon_i} = \hat{Y}_i(1-\hat{Y}_i) \ .$$

$\hat{Y}_i$ kan emellertid som nämnts bli <0 och >1 vilket ger upphov till negativa variansskattningar. Vi kan inte heller för sådana fall sätta $\hat{Y}_i$ till 0 respektive 1 vilket ger variansskattningen 0.

De möjligheter som står till buds är att antingen utesluta sådana observationer från analysen eller att sätta $\hat{Y}_i$ till något tal godtyckligt nära 0 respektive 1 såsom 0.01 respektive 0.99. I ett andra steg erhålles sedan en vägd minstakvadratskattning av $\alpha$ och $\beta$. Denna är visserligen effektiv i klassen av linjära estimatorer men övriga svagheter hos ovägd minstakvadratskattning återfinnes även hos den vägda, bl.a. att estimatorn kan anta värden utanför intervallet [0;1].

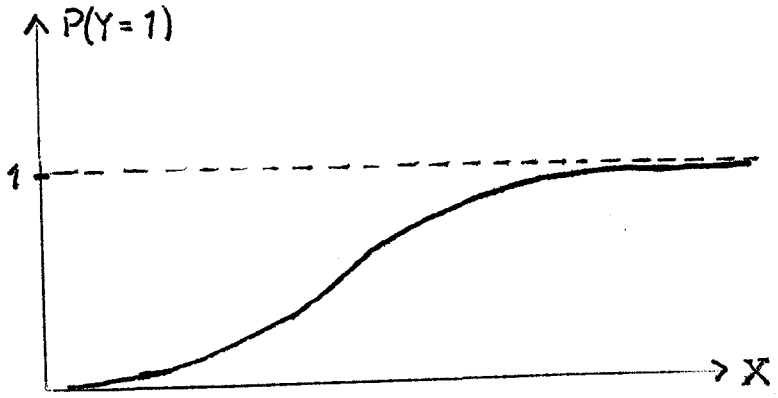
En ytterligare nackdel med den linjära regressionsmodellen är att normalfördelningsteorin ej fungerar för konstruktion av test och kondidensintervall på grund av att responsvariabeln är binär.

Den allvarligaste nackdelen med den linjära sannolikhetsmodellen är dock att vi i många fall a priori kan vara ganska övertygade om att den är felspecificerad. Det är inte rimligt att tänka sig att en ökning av X med en enhet skulle ha samma effekt oavsett från vilken nivå X ökar. Då X ökar från en nivå vid vilken sannolikheten att Y antar värdet 1 redan är hög (säg 0.95 eller högre) förväntar vi oss ingen större ökning av P(Y=1). Om X däremot ökar lika mycket från en nivå vid vilken sannolikheten P(Y=1) ligger i trakten av 0.5 förväntar vi oss en mycket större ökning av sannolikheten. Det är i de flesta tillämpningar inte heller troligt att Y=1 är en omöjlig händelse som inte kan inträffa om X är mindre än något tal a och att Y=1 är en säker händelse som måste inträffa om X är större än något tal b. (I resonemanget ovan har för enkelhets skull förutsatts att P(Y=1) ökar med X men naturligtvis kan P(Y=1) lika väl minska med ökande X-värden.) Slutsatsen blir att en sannolikhetsmodell som förutsätter att X:s effekt på P(Y=1) är av den principiella typ som beskrivs av figur 7.2 verkar mera adekvat än den linjära modellen.

Är då den linjära sannolikhetsmodellen helt utan värde ? Nej, det är den naturligtvis inte. Enkelheten hos den linjära modellen har ett värde som gör den tilltalande om den ger en någorlunda adekvat beskrivning. Om andelen individer som har värdet Y=1 inte är alltför låg eller alltför hög i

någon delgrupp i populationen är den linjära modellen ett konkurrenskraftigt alternativ. Om däremot vissa delgrupper har mycket låg frekvens (säg <10 %) eller mycket hög frekvens (säg >90 %) av värdet Y=1 är modellen i allmänhet helt orealistisk.

Figur 7.2. S-formad sannolikhetsmodell.




## 7.2 Logitmodellen

Det förekommer olika modeller vars allmänna drag är av den art som illustreras av principfiguren 7.2. Eftersom fördelningsfunktionen för en normalfördelad variabel har ett utseende av detta slag och med ett funktionsvärde begränsat till intervallet (0;1) förekommer det att man anpassar en sådan funktion för att bekskriva hur P(Y=1) beror av X. Man talar då om en probitmodell.
Numeriskt ganska snarlika resultat erhålles med logitmodellen, även benämnd modellen för logistisk regression. Man utgår här från oddset för att Y=1 (d.v.s. från P(Y=1)/P(Y=0) som anger sannolikheten att händelsen inträffar i förhållande till sannolikheten att händelsen ej inträffar) och antar en konstant relativ förändring av oddset då X-värdet ökar en enhet. Detta innebär att logaritmen för oddset blir en linjär funktion av X.

$$\log \frac{P_i}{1-P_i} = \alpha+\beta_i \qquad (7.4)$$

Det är vänsterledet d.v.s. logaritmen för oddset som benämns logit.

Logitmodellen kan också skrivas på formen

$$P_i = \frac{1}{1+e^{-(\alpha+\beta X)}} \qquad (7.5)$$

Ett positivt β-värde innebär att P ökar med ökande X medan ett negativt β innebär att P minskar med ökande X. Skillnaden mellan probit- och logitmodellerna ligger främst i att logitfördelningen har längre "svansar" än probitfördelningen vilket innebär att funktionen långsammare närmar sig värdena 0 respektive 1 för extrema X-värden. Det saknas enkel tolkning av parametern β. (En enhets ökning av X medför en relativ förändring av oddset för $P_i$ med faktorn $e^{\beta}$. Detta är emellertid inte särskilt intressant eftersom vi är vana vid att tänka i termer av sannolikheter men inte i termer av odds.)

## 7.3 Multipla logitmodeller

Logitmodellen kan på samma sätt som den vanliga enkla linjära regressionsmodellen utvidgas till att inkludera flera regressorer $X_1,\ldots,X_k$. Man erhåller

$$\log \frac{P}{1-P} = \alpha+\beta_1 X_1+\ldots+\beta_k X_k \qquad (7.5)$$

och ekvivalent

$$P = \frac{1}{1+e^{-(\alpha+\beta_1 X_1+\ldots+\beta_k X_k)}} \qquad (7.6)$$

Effekterna på logit d.v.s. på log P/(1-P) av förändringar av X-variablerna är additiva. Relativa förändringar av oddset P/(1-P) vid en ökning av variabeln $X_j$ med en enhet är således oberoende av utgångsvärdet för $X_j$ och oberoende av nivån för övriga X-variabler. Effekten på sannolikheten $P_i$ av en enhetsökning av variabeln $X_j$ framgår om vi deriverar P med avseende på $X_j$. Man erhåller (se Hanushek and Jackson [11] sid.188)

$$\frac{\partial P}{\partial X_j} = \beta_j P(1-P)$$

Eftersom P är en funktion av samtliga variabler $X_1,\ldots,X_k$ enligt (7.6) beror absoluta förändringen av sannolikheten P,

orsakad av en viss ökning av variabeln $X_j$, på nivån för $X_j$ samt på nivån för samtliga övriga X-variabler. Det finns således en interaktion mellan regressorer i dess inverkan på sannolikheten P.

Logitmodellen kan byggas ut med produkttermer av typen $X_iX_j$ så att man får en interaktion mellan regressorer även i dess inverkan på logit P d.v.s. på log P(1-P). De relativa förändringarna av oddset P/(1-P) vid en enhetsökning av variabeln $X_j$ blir då beroende av nivån på $X_j$ och av nivån på de övriga X-variabler som $X_j$ samspelar med.

## 7.4 Regressorstyper och skattningsmetoder för logistisk regression

Regressorerna kan bestå av kvantitativa variabler, kategoriska variabler eller båda dessa variabeltyper samtidigt. En kategorisk variabel kan ange en genuin kvalitativ egenskap eller representera en klassindelning av en kvantitativ variabel. De skattningsmetoder som kommer i fråga är vägd minstakvadratskattning och maximumlikelihoodskattning. Den förra förutsätter att observationerna kan sammanställas i en kontingenstabell med flera observationer i varje cell. En allmän diskussion av skattningsproblematiken återfinnes i Hanushek and Jackson kap. 7 [11].

## 7.5 Datorprogram för logistisk regression

Program för logistisk multipel regression finns bl.a. i BMDP (infört i manualen från 1979) och i SAS (infört i SAS Supplementary Library User's Guide 1980). Programmet har utnyttjats i kap. 8 för analys av politisk aktivitet.

# 8. REGRESSIONSANALYS I KOMMUNUNDERSÖKNINGARNA

## 8.1 Ett exempel: Politisk aktivitet

I kommunundersökningarna studeras i stort sett genomgående samband mellan kategoriska variabler. I de 10 rapporter som utkommit då detta kapitel skrives baseras i stor utsträckning resonemang och slutsatser på korstabelleringar av den beroende variabeln mot en eller två förklaringsvariabler i taget. I den mån man tillgriper statistiska analysmetoder dominerar helt regressionsanalysen och variansanalysen: Med enstaka undantag saknas helt beskrivningar av modeller och modellarbete. I många fall är detta säkerligen ett uttryck för att modelltänkandet inskränkt sig till frågan vilka förklaringsvariabler som skall ingå i regressionsmodellen eller variansmodellen.

I fråga om regressionsmodeller förekommer bara den enklaste typen av enekvationsmodeller som är linjär i både variabler och parametrar samt har konstant varians.

Endast i en rapport förekommer det en beskrivning av hur regressionsanalysen används. I "Medborgarna och kommunen" [29] studerar i fjärde kapitlet Westerståhl och Johansson medborgarnas politiska aktivitet och organisationsaktivitet. Modellerna anges inte explicit men man får ändå en översiktlig beskrivning av analysgången. Vi skall därför välja denna studie som utgångspunkt för en diskussion av regressionsanalysens användning.

Fram till sidan 78 ges en framställning som utnyttjar korstabelleringen för att beskriva hur aktivitetsgrad varierar med personegenskaper, attityder och kommunegenskaper. Kapitlet avslutas med en sammanfattning, varvid man tillgriper regressionsanalysen som analysinstrument. Denna sammanfattning återges i sin helhet.

> "Sammanfattande analys
>
> Vi har i detta kapitel sökt efter egenskaper som äger samband med den medborgerliga aktiviteten och som eventuellt kan förklara varför denna tar sig växlande uttrycksformer, ökar eller minskar. I vissa fall har det varit naturligt att förutsätta en

direkt orsak-verkan relation: missnöje med den kommunala servicen, exempelvis, bör stimulera politisk aktivitet. Men i många fall är sambanden mellan oberoende och den beroende variabeln, aktiviteten, mycket invecklade. Mellanliggande variabler och alla möjliga andra variabler kommer in i bilden och man får nöja sig med att konstatera en samvariation.

Vad vi nu skall göra är att låta dessa olika förklaringsvariabler än en gång passera revy. Den mest ambitiösa uppgiften, att pröva teoretiska modeller för att söka fastlägga de olika variablernas ordningsföljd och inbördes relationer, skall vi inte ge oss in på. Vi nöjer oss i stället med att föra in grupper av variabler i olika steg för att se vilket bidrag till förklaringen av de medborgerliga aktiviteternas variation som dessa variabler var för sig och alla tillsammans kan ge.
Framställningen begränsas till de tre slagen av politisk aktivitet och till organisationsaktivitet. Vidare gäller att vi av tekniska skäl måste frångå den hierarkiska skalan och åter behandla varje aktivitet för sig, d.v.s. utan hänsyn till att samma individer kan delta i olika slag av aktiviteter. Tre mått på förklaringsvariablernas effekt används: betakoefficienter (standardiserade partiella regressionskoefficienter) för att mäta den effekt som varje enskild variabel, med kontroll för effekten av övriga, har på det berörda slaget av aktivitet, en multipel korrelationskoefficient (R) för att mäta den samlade effekten hos alla förklaringsvariabler samt ett annat mått ($R^2$) som mäter samma sak i termer av "förklarad varians", varvid de angivna värdena direkt kan översättas till procentandel förklarad varians. Analysen framgår i tre steg.
Steg 1 (tabell 4:9) innefattar personliga bakgrundsegenskaper, inklusive politiskt engagemang i familjen. Den partipolitiska aktiviteten påverkas i första hand av ålder. Betakoefficienten .08 står för den effekt i riktning mot ökad politisk aktivitet som steget från yngre till medelålders har och .14 för den starkare effekt som steget från yngre till äldre har. En lika stor effekt har det förhållandet att det i familjen finns någon politiskt engagerad person.
Kön, d.v.s. det förhållandet att vederbörande är man och inte kvinna, ger en betakoefficient på .11. Även utbildning och socialgrupp ger vissa mindre effekter. Politisk aktivitet via tjänstemän bestäms däremot främst av de mer renodlade "statusvariablerna" utbildning och socialgrupp. För att sannolikheten att utnyttja denna kanal skall bli så stor som möjligt, skall man vara högt utbildad och tillhöra socialgrupp 1. Kön spelar här ingen som helst roll. Vad beträffar politisk aktivitet genom aktioner har de olika bakgrundsegenska-

perna i allmänhet något mindre betydelse. Det framgår att kön inte heller här spelar någon roll. Om man tillhör den äldsta åldersgruppen, har enbart obligatorisk utbildning eller tillhör socialgrupp III, är sannolikheten att utöva denna typ av politisk aktivitet mindre än i övriga fall.
Ifråga om organisationsaktivitet förändras däremot bilden. Här blir kön den mest väsentliga faktorn. Är man dessutom medelålders och, liksom när det gällde aktivitet via tjänstemän, högutbildad och tillhör socialgrupp 1 ökar sannolikheten för organisationsaktivitet. Här möter man alltså hela uppsättningen av traditionella "elitegenskaper".
I steg II tillkommer kommuntyp, kravbredd och syn på service som förklaringsvariabler. När vi nu alltså kontrollerar för kravbredd, ökar effekten på det partipolitiska deltagandet av att tillhöra den äldre åldersgruppen. Tolkningen bör vara att de äldre har mindre krav men har högre sannolikhet att delta än en yngre person med lika många krav. Kravbredden är en ny väsentlig egenskap när det gäller politiskt deltagande. Också i frågan om aktivitet via tjänstemän har kravbredd en viss betydelse. Flera krav ökar sannolikheten att vara aktiv. Effekten av hög utbildning och att tillhöra socialgrupp 1 påverkas däremot inte. Det finns alltså inget samband mellan sistnämnda egenskaper och kravbreddden.
Också direkt aktivitet påverkas av kravbredden, men mest värt att notera är effekten av kommuntyp. Detta är det enda slag av aktivitet för vilken egenskapen att bo i en viss typ av kommun är mer väsentlig än någon annan av de egenskaper som här analyserats. Hög kravbredd ökar också sannolikheten för organisationsaktivitet. Även i detta fall är effekten av övriga egenskaper oförändrad, d.v.s. kravbredden har inget samband med dessa egenskaper.
I steg III tillföres ideologifaktorerna och egenskapen personlig kompetens. Ideologifaktorerna har begränsad effekt: kommunal utjämning som inte har någon signifikant effekt alls har uteslutits ur redovisningen. I och med att egenskapen personlig kompetens införes påverkas de tidigare betavärdena kraftigt (det finns inget skäl att tro att ideologifaktorerna har denna effekt). Obestridligen är såväl kön som ålder, utbildning, socialgrupp och politiskt engagemang under uppväxten egenskaper som påverkar den personliga kompetensen. Denna har i sin tur stor betydelse för partipolitisk aktivitet. Det finns uppenbarligen också ett avsevärt samband mellan kravbredd och personlig kompetens. Hög personlig kompetens är så kraftigt korrelerad med stor kravbredd att den förstnämnda helt tar överhanden (det verkar inte riktigt att se personlig kompetens som en mellanliggande variabel här).

Också aktivitet via tjänstemän påverkas av personlig kompetens, om än inte i lika stor utsträckning. Hög utbildning och tillhörighet till socialgrupp 1 har kvar mera av sin betydelse. Även för de aktionsaktiva har den personliga kompetensen betydelse. Denna egenskap blir viktigare än det faktum att man bor i en viss kommuntyp. Eftersom dessa egenskaper saknar samband, behåller dock kommuntyp sitt betavärde (i stort sett). Däremot tar den personliga kompetensen upp en stor del av kravbredd och tillhörighet till socialgrupp II. Organisationsaktivitet bestäms också till stor del av politisk kompetens. Vi får samma effekt som tidigare i form av lägre betavärden för främst högre utbildning och kravbredd.
Den samlade effekten av dessa förklaringsvariabler - som högst 19% förklarad varians - kan förefalla låg, inte minst med tanke på de tydliga samband som den tidigare, omfattande tabellredovisningen har indikerat. Förklarad varians är ett krävande mått som ofta ger låga värden även när påtagliga och signifikanta samband föreligger. Det bör också påpekas att man i motsvarande undersökningar i andra länder nått ungefär samma förklaringsnivå (jfr Verba m.fl., 1978) så länge man begränsar sig till att söka förklara beteenden. Att förklara attityder med attityder är ofta lättare.
På steg I blev andelen förklarad varians för de tre slagen av politisk aktivitet av samma storleksordning (5-6%). Däremot skilde organisationsaktiviteten ut sig med en väsentligt högre andel förklarad varians, betingad av de traditionella personliga bakgrundsegenskapernas inverkan (14%). Tillskottet av förklarad varians genom steg II blev inte särskilt stort och relationerna mellan de olika aktivitetsformerna i huvudsak oförändrad. Här var det framför allt kravbredden som hade effekt liksom kommuntyp för aktionsaktiviteten. På det sista steget III inträdde en väsentlig förändring i första hand ifråga om den partipolitiska aktiviteten genom införande av variabeln personlig kompetens: andelen förklarad varians fördubblades i det närmaste (från 9 till 17%). Även övriga aktivitetsformer fick ett tillskott av förklarad varians. Det är givet att variabeln personlig kompetens som den här definierats ligger nära den partipolitiska aktiviteten. I vissa undersökningar har man lagt in ett dylikt psykologiskt mått i själva aktivitetsmåttet och då också ofta kunnat förklara en större del av den politiska aktivitetens varians. Vi har, som sagt, föredragit att så långt möjligt endast använda aktivitetsmått som mäter beteende.
I en slutsammanfattning kan vi konstatera att personlig kompetens (innefattande lokalt uppmärksamhetsindex, åsiktsbredd, intresse för politik, självuppskattad kännedom om politik, deltagande i diskussion om kommu-

Tabell 8.1. Utdrag från Westerståhl och Johansson [29] sid. 80.

**Tabell 4:9 Fyra slag av medborgerlig aktivitet – den samlade effekten av olika förklaringsvariabler (Betakoefficienter, multipel korrelation, förklarad varians)**

| STBG | | Män | Medelålders | Äldre | Medelutbildn. | Högutbildn. | Soc.grupp II | Soc.grupp I | Pol.engagemang i familjen | Kommuntyp (Största eff.) | Servicesyn | Kravbredd | Närdemokrati | Expertstyre | Systemgillande | Personlig kompetens | R | $R^2$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Aktivitet: | | | | | | | | | | | | | | | | | |
| | genom partier | .11 | .08 | .14 | .04 | .08 | .05 | .08 | .14 | | | | | | | | .25 | .06 |
| | tjänstemän | * | .05 | -.05 | * | .12 | * | .12 | .07 | | | | | | | | .25 | .06 |
| I | aktioner | * | * | -.10 | .08 | .07 | .09 | .07 | .04 | | | | | | | | .21 | .05 |
| | organisationer | .21 | .12 | -.07 | .09 | .13 | .07 | .10 | .08 | | | | | | | | .38 | .14 |
| | partier | .12 | .08 | .18 | * | .08 | .05 | .08 | .14 | .06 | .04 | .11 | | | | | .29 | .09 |
| | tjänstemän | * | .06 | * | * | .11 | .04 | .12 | .06 | * | .06 | .08 | | | | | .28 | .08 |
| II | aktioner | * | * | -.07 | .07 | .06 | .09 | .07 | .04 | .11 | .07 | .08 | | | | | .27 | .07 |
| | organisationer | .22 | .12 | * | .08 | .11 | .07 | .09 | .08 | * | * | .12 | | | | | .40 | .16 |
| | partier | .06 | * | .12 | * | * | * | .05 | .09 | * | .04 | * | -.06 | * | .08 | .28 | .41 | .17 |
| | tjänstemän | * | .05 | * | * | .08 | * | .10 | .04 | * | .05 | .05 | * | * | * | .17 | .32 | .10 |
| III | aktioner | * | * | -.07 | .06 | * | .07 | .05 | * | .10 | .05 | * | * | -.05 | * | .16 | .30 | .09 |
| | organisationer | .17 | .10 | -.08 | .06 | .07 | .06 | .07 | .05 | .05 | * | .07 | -.07 | * | .04 | .19 | .43 | .19 |

* ej signifikant på 95 %-nivån

nala frågor) har den största förklaringskraften när det gäller partipolitisk aktivitet men också mycket stor effekt på de andra formerna av aktivitet. Eftersom denna egenskap ligger så nära den partipolitiska aktiviteten, är det kanske av större intresse att se på de olika förklaringsvariablernas effekt före steg III.
Den partipolitiska aktiviteten påverkas då i första hand av hög ålder, politiskt engagemang i familjen, manligt kön och kravbredd. För dem som är aktiva via tjänstemän är det framför allt hög utbildning samt tillhörighet till socialgrupp I som spelar roll. Vad aktionsaktiviteten beträffar ger som nämnts kommuntillhörigheten det starkaste utslaget, närmast det förhållandet att man bor i stora kommuner. Organisationsaktiviteten påverkas slutligen främst av kön, ålder, kravbredd och hög utbildning.
De olika slagen av medborgerlig aktivitet uppvisar alltså skilda sambandsmönster, de påverkas av skilda förhållanden i samhället. Om man förutsätter att aktivitet ger inflytande, säger dessa samband också något om vilka som utövar inflytande på kommunens verksamhet. De säger även något om vad som händer om en väg för inflytande ersättes av andra."

## 8.2 Modellvalet och analysen

Det är belysande att diskutera den genomförda analysen ur en rad olika aspekter. Vi skall här beröra

- kommunvariabler
- tänkbara orsakssamband mellan variablerna
- variabelkonstruktioner
- binär responsvariabel
- urvalsförfarandet
- den genomförda analysen
- alternativa modellansatser.

## 8.3 Kommunvariabler

De modeller som konstrueras avser individnivå. Modellerna skall om möjligt förklara hur sannolikheten för politisk aktivitet varierar med vissa bakgrundsvariabler som karakteriserar individen eller dennes situation. Invånarantalet i kommunen torde vara ett uttryck för den miljö i vilken den enskildes eventuella politiska aktivitet äger rum. Med

ökande folkmängd ökar storlek och komplexitetsgrad hos politisk apparat och förvaltningsapparat. Den enskilde individens situation torde präglas av huruvida denne är bosatt i tätort eller glesbygd och i det senare fallet även av avståndet till tätort. Tätortsgradens värde som förklaringsvariabel synes därför tveksamt.
En annan faktor av betydelse torde vara hur väl det politiska livet fungerar i kommunen och hur väl utbyggd samhällets service är inom kommunen. Författarna har inte haft tillgång till objektiva data om dessa förhållanden. Däremot utnyttjas data om invånarnas åsikter och attityder i dessa avseenden. Konsekvensen av att inkludera attitydvariabler diskuteras i kap. 8.4 och 8.5.

## 8.4 Kausala relationer

Författarna försöker förklara förekomst eller avsaknad av en viss politisk aktivitet hos individen med hjälp av bakgrundsegenskaper hos individen, egenskaper hos kommunen och attityder hos individen. Denna omfattar

kön
ålder
utbildning
socialgrupp
politiskt engagemang i familjen under uppväxttiden
kommuntyp (invånarantal och tätortsgrad)
bedömning av och anspråk på kommunal service
ideologisyn beträffande kommunal demokrati
"kompetens" för att inneha kommunalt uppdrag.

Alla egenskaper (utom politiskt engagemang i familjen) avsåg undersökningstillfällets förhållanden.
Som allmänt mål vid modellbyggnad gäller att man vill konstruera en modell som är så enkel som möjligt men ändå kan ge en adekvat beskrivning av sakförhållandena. Det engelska uttrycket "a parsimonious model", som saknar bra svensk översättning, innebär en modell som är sparsam med variabler och parametrar. I samhällsvetenskapliga tillämpningar kan man aldrig konstru-

era modeller som innehåller alla variabler som i verkligheten har ett inflytande. Man måste på förhand utesluta variabler som man är säker på har relativt marginella effekter. Likaledes bör man också i allmänhet undvika att inkludera två eller flera variabler som speglar i stort sett samma förhållanden. I det aktuella fallet kan socialgruppstillhörigheten tänkas vara starkt beroende av utbildningsnivån, varför det borde prövats om inte endera av dessa förklaringsfaktorer kunnat utgå.

I modellen som valts ingår en rad variabler som speglar inställningen till utbudet av kommunal service, ideologisyn och personlig kompetens vars orsak-verkanrelation till politisk aktivitet är synnerligen oklar. Troligen föreligger någon form av växelverkan mellan politisk aktivitet och attitydvariablerna. Attityder är föränderliga i tiden. Den kommunala servicen förändras också och därmed rimligen i viss utsträckning medborgarnas syn på servicen. Det är därför svårt att se dagens attityder som lämpliga förklaringsfaktorer till en politisk aktivitet som ofta pågått under längre tid och kanske inletts för decennier sedan.

En enligt min mening närliggande konsekvens borde vara att man försöker beskriva variationer i den politiska aktiviteten med de renodlade bakgrundsvariablerna kön, ålder, utbildning (alternativt socialgrupp), politiskt engagemang i familjen under uppväxttiden och kommuntyp. (Synpunkter på kommunvariablernas lämplighet ges i kap. 8.12.) Syftet skulle då närmast betraktas som prediktivt trots att det i botten ligger en önskan att förklara vilka faktorer som påverkar den politiska aktiviteten. I annat fall räcker det inte med en enkel enekvationsmodell med politisk aktivitet som en funktion av ett antal faktorer. Man måste i så fall, exempelvis med hjälp av ett ekvationssystem, specificera beroendeförhållandena mellan varje "responsvariabel" och övriga variabler. Detta leder till ett komplext modellbyggnadsarbete och kräver, förutom grundlig sakkännedom, djupa kunskaper i statistik.

## 8.5 Variabelkonstruktioner

Variablerna som ingår i analysen finns inte explicit definierade. De är delvis möjliga att rekonstruera ur Tabell 4.9 och ur tidigare framställning i rapporten. De beroende variablerna är behandlade var för sig och definieras med här införda beteckningar enligt

$$A_1 = \begin{cases} 1 & \text{om personen är politiskt aktiv genom partier} \\ 0 & \text{eljest} \end{cases}$$

Variablerna $A_2$, $A_3$ och $A_4$ anger på motsvarande sätt aktivitet genom tjänstemän, aktivitet genom aktion (undertecknande av upprop, deltagande i demonstration) respektive aktivitet genom organisationer.

Vid kontakt med författarna befanns att förklaringsvariablerna var definierade som följer

$$KV = \begin{cases} 1 & \text{för kvinnor} \\ 0 & \text{för män} \end{cases}$$

Åldern behandlas som en kvalitativ variabel medelst indelning i tre åldersklasser vilka särskiljes medelst dummyvariabler (indikatorvariabler).

$$ÅLDH = \begin{cases} 1 & \text{för åldern > 60 år} \\ 0 & \text{eljest} \end{cases}$$

$$ÅLDM = \begin{cases} 1 & \text{för åldern 30-60 år} \\ 0 & \text{eljest} \end{cases}$$

(H och M står här för hög respektive medel.)

Utbildning delas i tre klasser (enligt sid 35).

$$UTBH = \begin{cases} 1 & \text{för högre utbildning} \\ 0 & \text{eljest} \end{cases}$$

$$UTBM = \begin{cases} 1 & \text{flerårig yrkesutbildning eller teoretisk utbildning högst 2 år} \\ 0 & \text{eljest} \end{cases}$$

Socialgruppsindelningen ges av

$$SOC1 = \begin{cases} 1 & \text{för socialgrupp I} \\ 0 & \text{eljest} \end{cases}$$

$$SOC2 = \begin{cases} 1 & \text{för socialgrupp II} \\ 0 & \text{eljest} \end{cases}$$

Variabeln Politiskt engagemang i familjen avser fråga 49 i bilaga 1: "När Ni växte upp, var det då någon i familjen eller i Er närmsta omgivning som hade uppdrag i någon kommunal nämnd eller i fullmäktige ?" (Svarsalternativ: ja, nej, minns ej/vet ej)

$$POL = \begin{cases} 1 & \text{för svar "ja" på fråga 49} \\ 0 & \text{för svar "nej"} \end{cases}$$

Svaret "minns ej/vet ej" har betraktats som "missing data". Personer som givit detta svar har uteslutits vid analysen. Motsvarande gäller övriga frågor.
Kommuntyp ger en indelning i 9 klasser efter invånarantal och tätortsgrad. 8 dummyvariabler $K_1,\ldots,K_8$ används för klassificering.
Variabeln Servicesyn bestäms av svaren på fråga 6:

> "Vi har talat om kommunens insatser på olika serviceområden. Hur skulle Ni på det hela taget vilja bedöma den kommunala servicen här i Er kommun ? Vilket av svaren på de här korten stämmer bäst med Er uppfattning ?"

Visa svarskort 6

1. Jag är helt nöjd med kommunens service.
2. Jag är i stort sett nöjd med kommunens service.
3. Jag är ganska nöjd med kommunens service, men tycker att den kunde varit bättre i vissa fall.
4. Jag är övervägande missnöjd med kommunens service, men tycker att den är bra i vissa fall.
5. Jag är helt missnöjd med kommunens service.
7. Ingen åsikt/vet ej.

Svar 7 har betraktats som svarsbortfall och individer som avgivit detta svar ingår ej i analysen.
Servicesyn har behandlats som en intervallskalevariabel vars värden utgöres av svarskoderna 1 - 5.
Variabeln Kravbredd anger antal områden bland 14 angivna inom vilka respondenten anser att kommunen borde göra mera (Fråga 5 A.).
Variablerna Närdemokrati, Expertstyre och Systemgillande anger så kallade faktorscores, erhållna vid en faktoranalys av ett större antal enkla frågor avsedda att belysa medborgarnas ideologisyn beträffande kommunal demokrati (se Westerståhl och Johansson sid 35-38).
Variabeln Personlig kompetens slutligen beskrivs på följande sätt (sid 59).

> "Personlig kompetens har mätts som ett additivt index byggt på följande moment: intresse för lokalt material i media, åsiktsbredd, intresse för politik, kännedom om kommunalpolitik och deltagande i diskussioner om kommunala frågor.
> Indexet är konstruerat på följande sätt:
>
> 1. Lokalt uppmärksamhetsindex (Lennart Brantgärde)
>    a) läsning om kommunalpolitik i dagstidningar
>    b) läsning om lokala nyheter i dagstidningar
>    c) ser på regional-TV
>    d) hör på lokalradio.
> 2. Åsiktsbredd. Antal frågor i vilka man uttalat en åsikt (29 möjliga).
> 3. Intresse för politik (4 steg, jfr fråga 38 i bilaga 1)
> 4. Självuppskattad kännedom om politik (4 steg, jfr fråga 33 i bilaga 1).
> 5. Deltagande i diskussion om kommunala frågor (3 steg, jfr fråga 37 i bilaga 1).
>
> Under punkt 1 ingår moment a-d med lika vikt. Indexet på personlig kompetens är därefter konstruerat så att punkterna 1-5 ingår med lika vikt och ger till resultat en kontinuerlig variabel som varierar mellan 0 och 100. Ett högt värde på detta index anger hög personlig kompetens".

Variablerna Servicesyn, Närdemokrati, Expertstyre, Systemgillande och Personlig kompetens har behandlats som intervallskalevariabler av författarna.

Svarsalternativen på servicefrågan har en klar rangordning vilken skall belysas av åsatta kodvärden. Men därutöver kan dessa väljas helt godtyckligt. Kravbredd är en antalsvariabel. Variablerna Närdemokrati, Expertstyre och Systemgillande har uppkommit genom transformationer (faktoranalys) av kvalitativa observerade variabler. Personlig kompetens är ånyo en helt godtycklig skalkonstruktion. Det är därför bortsett från Kravbredd helt oacceptabelt att behandla dessa variabler som intervallskalevariabler.

## 8.6 Tillåtna skaltyper för regressorerna

Regressorerna i regressionsmodeller kan vara av två slag: kvantitativa variabler av intervall- eller kvotskaletyp och indikatorvariabler (dummyvariabler) som anger huruvida individen tillhör eller icke tillhör en viss kategori (har eller saknar en viss egenskap).
Variabelkonstruktioner i vilka variabelvärdena är godtyckliga eller enbart skall spegla en viss rangordning är föga meningsfulla som regressorer, eftersom parametervärdena i modellen starkt påverkas av skalkonstruktionen. I princip kan man ge en sådan regressor vilka regressionseffekter som helst genom att välja variabelvärden på lämpligt sätt.
Indexet Personlig kompetens är uppbyggt av ett antal delindex (se beskrivning ovan). Skalorna som mäter enskilda egenskaper har rangordning. De är även konstruerade med lika skalsteg vilket emellertid inte ger någon ytterligare information. Inte ens om om man konstruerar ett additivt totalindex under förutsättning att en enskild skala skall ha lika skalsteg kan man ge detta index rangordningsegenskaper. En entydig rangordning är omöjlig att åstadkomma av två skäl.

> Ett delindex med exempelvis skalvärdena 1,2,3,4 är fullständigt ekvivalent med varje multipel av dessa värden, såsom 2,4,6,8 eller 10,20,30,40. Olika val av skalsteg medför att rangordningen av individerna enligt totalindexet kan bli olika.
>
> Delindex kan vägas ihop med olika viktsystem vilka kan ge helt skilda rangordningar av individerna enligt totalindexet.

Den normering av totalindexet till variationsområdet 0-100 som författarna gjort avhjälper på intet sätt totalindexets brister.

## 8.7 Binär responsvariabel

Responsvariabeln är binär med värden 1 (politiskt aktiv) och 0 (ej politiskt aktiv). Den vanliga linjära regressionsmodellen är, som framgår av kap. 7, av olika skäl ofta mindre lämplig i sådana situationer. Den genomsnittliga frekvensen aktiva (A=1) i urvalet är 17%. För vissa delgrupper ligger andelen aktiva under 5%. Det är just i situationer då det finns delgrupper för vilka sannolikheten P ligger mycket lågt eller mycket högt som den linjära sannolikhetsmodellen i allmänhet måste anses som orealistisk (se kap. 7.1).

## 8.8 Hur bra modell är möjlig ?

För att en regressionsmodell med en binär responsvariabel skall uppvisa en någorlunda hög determination, till exempel i termer av förklarad varians, krävs att det med hjälp av vissa förklaringsvariabler är möjligt att avgränsa en grupp personer inom vilken de allra flesta har den aktuella responsegenskapen (Y=1) medan bland övriga personer de allra flesta saknar denna egenskap (Y=0).
Det säger sig själv att denna situation sällan påträffas. Man kan därför som framhållits i kap. 6.10 förvänta sig att modeller för binära responsvariabler i allmänhet får en ytterst låg determinationsgrad.
Man skall inte fästa särskilt stort avseende vid determinationen vid analys av binära responsvariabler. Determinationen mäter inte vad man primärt är intresserad av nämligen avvikelse mellan observerad andel positiva svar (f) och predikterad sannolikhet ($\hat{P}$) för positivt svar inom olika delgrupper. Ett enkelt mått på detta ges av genomsnittlig absolut avvikelse vanligen betecknad MAD efter den engelska termen "mean absolute deviation"

$$MAD = \frac{1}{m} \Sigma |f - \hat{P}|$$

där m anger antal icke tomma celler i kontingenstabellen med uppdelning efter förklaringsvariablerna. MAD tar icke hänsyn till att osäkerheten i skattningen av $\hat{P}$ varierar mellan olika celler.

Westerståhl och Johansson använder standardiserade regressionskoefficienter (betakoefficienter) för att ange de olika förklaringsvariablernas betydelse i dess inverkan på sannolikheten för politisk aktivitet. Detta är emellertid olyckligt eftersom det inte är meningsfullt att tala om en standardavvikelses ökning av en förklaringsvariabel som bara kan anta värden 0 och 1. (Se kap. 6.16). Den vanliga regressionskoefficienten ger däremot ett lätt tolkbart mått, nämligen förväntad förändring av sannolikheten för politisk aktivitet om man jämför personer utan den speciella bakgrundsegenskapen med personer som har denna egenskap (förutsatt att personerna i övrigt har samma bakgrundsegenskaper).

## 8.9 Komplext urval

Urvalsfraktionen i medborgarstudien varierar starkt mellan olika kommuner. I 45 av de 50 urvalskommunerna utvaldes 40 personer från varje kommun. I de övriga 5 kommunerna (Sjöbo, Kävlinge, Grästorp, Lidköping och Luleå) som var föremål för specialstudier utvaldes 300 personer per kommun. Följden blev att Grästorp fick en urvalsfraktion som var cirka 200 gånger större än Uppsalas (de båda extremerna). Av hela bruttourvalet kommer således 1500 personer från de 5 specialstuderade kommunerna och övriga 1800 personer från övriga 45 kommuner.

För individegenskaper vars fördelningar inom olika kommuner är markant olika kan således ett ovägt och en vägd fördelning (vikt = inverterade värdet av nettourvalsfraktionen) för hela urvalet skilja sig klart åt. Bland de variabler som ingår vid analysen av politisk aktivitet illustreras detta mest markant av utbildningsvariabeln.

| | Utbildningsfördelning (%) | | |
|---|---|---|---|
| | Endast obligatorisk | Högst 2 år över oblig. | Mer än 2 år över oblig. |
| Ovägt | 52.2 | 21.1 | 26.6 |
| Vägt | 43.7 | 24.0 | 32.3 |

Av kommunaldemokratiska forskningsgruppens rapporter framgår i allmänhet inte om andelar för olika variabler är vägda eller ovägda. Det bör framhållas att vägningar ofta kan ha större betydelse vid studium av flerdimensionella fördelningar och samvariation än vid studium av fördelningen för en variabel.

Det finns fullständiga data för 2452 personer angående de presumtiva förklaringsvariablerna för politisk aktivitet (bortfallet är alarmerande högt men detta problem behandlas inte här).

För att illustrera skillnaderna mellan vägda och ovägda skattningar ges ovägda och vägda skattningar av elementen i korrelationsmatrisen för de presumtiva förklaringsvariablerna. (Tab. 8.2.) Variabeldefinitionerna framgår av kap. 8.4. Här skiljs emellertid endast på socialgrupp 1 (SOC=1) och socialgrupp 2+3 (SOC=0) och på obligatorisk utbildning (UTB=0) och utbildning över den obligatoriska (UTB=1). Följande variabler tillkommer utöver tidigare definierade.

| | | | | |
|---|---|---|---|---|
| TÄTM | 50% | < | andel tätortsbef. | ≤ 80% |
| TÄTH | | | andel tätortsbef. | > 80% |
| INVM | 11000 | < | andel invånare | ≤ 27000 |
| INVH | | | antal invånare | > 27000 |

Som synes av tab. 8.2 förekommer mycket stora skillnader mellan ovägda och vägda skattningar vad beträffar inbördes korrelationer mellan kommunvariabler samt mellan kommunvariabeln TÄTM och individvariabeln UTB. Ovägda skattningar av regressionsmodeller skulle därför komma att skilja sig ganska mycket från vägda skattningar. Detta gäller inte bara effekten av kommunvariabler och utbildning.

Om hänsyn ej tas till urvalssannolikheter vid estimationen kommer resultaten att i mycket hög utsträckning präglas av

datamönstren i de 5 specialstuderade kommunerna Sjöbo, Kävlinge, Grästorp, Lidköping och Luleå, som ensamma svarar för nästan lika många observationer som övriga 45 kommuner tillsammans.

Tabell 8.2 Korrelationsmatris för förklaringsvariabler. Rad 1 anger ovägd korrelation, rad 2 anger vägd korrelation. Största skillnader inringade.

| | INV2 | INV3 | TÄTM | TÄTH | KÖN | ÅLDM | ÅLDH | SOC | UTB | POL |
|---|---|---|---|---|---|---|---|---|---|---|
| INV2 | 1.00 | -0.58 | 0.16 | -0.13 | 0.01 | 0.00 | 0.01 | -0.01 | -0.03 | -0.01 |
| | 1.00 | -0.79 | 0.31 | -0.29 | 0.03 | 0.00 | 0.03 | -0.04 | -0.12 | -0.03 |
| INV3 | | 1.00 | -0.08 | 0.49 | -0.01 | 0.03 | -0.06 | 0.05 | 0.14 | 0.01 |
| | | 1.00 | -0.41 | 0.53 | -0.04 | 0.02 | -0.05 | 0.05 | 0.17 | 0.02 |
| TÄTM | | | 1.00 | -0.59 | 0.04 | -0.02 | 0.02 | 0.03 | -0.01 | 0.03 |
| | | | 1.00 | -0.90 | 0.04 | -0.02 | 0.06 | -0.03 | -0.15 | 0.01 |
| TÄTH | | | | 1.00 | -0.03 | 0.04 | -0.08 | 0.03 | 0.16 | -0.01 |
| | | | | 1.00 | -0.05 | 0.02 | -0.07 | 0.04 | 0.18 | 0.00 |
| KÖN | | | | | 1.00 | -0.01 | -0.03 | -0.07 | -0.01 | -0.01 |
| | | | | | 1.00 | -0.01 | -0.02 | -0.10 | -0.01 | -0.03 |
| ÅLDM | | | | | | 1.00 | -0.49 | 0.07 | -0.01 | -0.01 |
| | | | | | | 1.00 | -0.46 | 0.08 | 0.01 | 0.06 |
| ÅLDH | | | | | | | 1.00 | -0.09 | -0.33 | 0.01 |
| | | | | | | | | -0.12 | -0.28 | -0.04 |
| SOC | | | | | | | | 1.00 | 0.26 | 0.03 |
| | | | | | | | | 1.00 | 0.24 | 0.03 |
| UTB | | | | | | | | | 1.00 | 0.12 |
| | | | | | | | | | | 0.12 |
| POL | | | | | | | | | | 1.00 |

## 8.10 Stegvis genomförande av regressionsanalysen

Tidigare i detta kapitel har starkt kritiska synpunkter framförts beträffande modellvalet mot bakgrund av ömsesidiga kausala beroenden mellan variablerna, skalkonstruktioner för regressorerna (av attitydtyp) och responsvariabelns binära form.

Trots att Westerståhl och Johanssons linjära multipla regressionsmodell således måste betraktas som olämplig är det av in-

tresse att diskutera analysgången. Författarna säger:

> "Den mest ambitiösa uppgiften att pröva teoretiska modeller för att söka fastlägga de olika variablernas ordningsföljd och inbördes relationer, skall vi inte ge oss in på. Vi nöjer oss i stället med att föra in grupper av variabler i olika steg för att se vilket bidrag till förklaringen av de medborgerliga aktiviteternas variation som dessa variabler var för sig och alla tillsammans kan ge."

I ett första steg skattar Westerståhl och Johansson en modell med enbart personliga bakgrundsegenskaper som regressorer (kön, ålder, utbildning, socialgrupp, politiskt engagemang i familjen).

Författarna studerar estimaten och drar bl.a. slutsatsen att åldern är den viktigaste förklaringsvariabeln. I andra steget införes även kommuntyp, kravbredd och syn på service som förklaringsvariabler. Resultatet tolkas på följande sätt.

> "När vi nu alltså kontrollerar för kravbredd, ökar effekten på det partipolitiska deltagandet av att tillhöra den äldre åldersgruppen. Tolkningen bör vara att de äldre har mindre krav men har högre sannolikhet att delta än yngre personer med lika många krav."

Tyvärr är detta sätt att tolka koefficientförändringar vid stegvisa modifikationer av modellen helt felaktigt. Både modellen i steg 1 och modellen i steg 2 är felaktiga eftersom författarna anser det nödvändigt att införa ytterligare variabler i ett tredje steg. Skattningarna av ålderseffekten i steg 1 och i steg 2 blir därför systematiska felskattningar eftersom modellerna är felspecificerade. Det systematiska felets storlek beror på ett komplicerat sätt på samvariationen mellan åldern och alla i modellen saknade relevanta förklaringsvariabler som är korrelerade med åldern (jämför kap. 5.3 avseende felaktigt väntevärdesantagande).

Visserligen arbetar man sig ofta fram till en slutgiltig modell genom att stegvis pröva och modifiera en utgångsmodell, men alla utom den slutliga modellen måste betraktas som felaktiga och det är inte meningsfullt att försöka tolka skattningar av dessa modeller.

## 8.11 Westerståhl och Johanssons slutmodell

Inför förutom beteckningarna i kap. 8.4 (där också alla variabler, inklusive nedanstående, är definierade) följande

SERV = servicesyn
KRAV = kravbredd
NÄRD = närdemokrati
EXP = expertstyre
SYST = systemgillande
KOMP = personlig kompetens

Westerståhl och Johanssons slutmodell ser ut på följande sätt skriven symboliskt

A = KV ÅLDH ÅLDM UTBH UTBM SOC1 SOC2 POL
K1 K2 K3 K4 K5 K6 K7 K8 SERV KRAV
NÄRD EXP SYST KOMP (8.1)

vilket innebär

$$A = \alpha + \beta_1 KV + \ldots + \beta_{22} KOMP + \varepsilon \qquad (8.2)$$

Tabell 4.9 visar att endast 8 av 22 marginella koefficienttest givits signifikant resultat på 5%-nivån för modellen som förklarar aktiviteten genom partier (översta raden för steg III). Då man betraktar alla fyra aktivitetstyperna (genom partier, genom tjänstemän, genom aktioner, genom organisationer) finner man att enbart två regressorer, nämligen socialgrupp och personlig kompetens, givit signifikans vid de marginella testen för samtliga aktivitetstyper.
Som tidigare framhållits verkar det rimligt att anta att det föreligger en ömsesidig påverkan mellan attityder och politisk aktivitet. Det krävs då flera ekvationer för att åstadkomma en adekvat förklaringsmodell. Attitydvariablerna är också på grund av skalkonstruktionen (se kap. 8.4) helt olämpliga i en regressionsmodell. Det verkar därför rimligt att konstruera en modell med enbart personliga bakgrundsvariabler och eventuellt kommunvariabler som regressorer.
En grundläggande fråga är då om man i likhet med Westerståhl och Johansson kan anta att effekterna av olika bakgrundsvari-

abler är additiva. Detta verkar föga troligt och den additiva modellen kan också ge upphov till negativa prediktioner för Y-variabeln för vissa grupper.

## 8.12 Prövning av additiv modell

Bortsett från de ovan nämnda problemen är antalet regressorer i W o J:s modell (22 stycken) synnerligen högt. (Antalet egenskaper som används för att förklara variationerna i aktivitetsgrad är betydligt färre men många dummyvariabler åtgår för kategoriindelningarna, främst avseende kommuntyp). På grund av kombinationen av många regressorer och korrelationer mellan regressorer, kanske främst de starka korrelationerna mellan dummyvariablerna för kommunindelningen, blir den slumpmässiga osäkerheten i parameterskattningarna mycket hög. I en sådan situation är det mycket svårt för att inte säga ogörligt att hitta en adekvat förklaringsmodell. Man måste realistiskt sett vara nöjd om man kan konstruera en god prediktionsmodell (se utförligare diskussion i kap. 6.1, 6.12 och 6.13).
Det är därför angeläget att undersöka om antalet regressorer kan minskas utan att man utelämnar någon individ- eller kommunegenskap som anses betydelsefull.
I stället för W o J:s indelning i 9 kommunklasser efter invånarantal och tätortsgrad med hjälp av 8 indikatorvariabler görs oberoende uppdelningar efter invånarantal och tätortsgrad. Klasserna är $\leq$ 11000, 11001-27000, > 27000 respektive $\leq$ 50%, 50% - 80%, > 80% vilka indikeras av de 4 indikatorvariablerna INVM, INVH respektive TÄTM, TÄTH (definierade i kap. 8.8).
För att pröva om en modell med additiva effekter för individ- och kommunvariabler är realistisk skattas MODELL 1:

A = INVM INVH TÄTM TÄTH KV ÅLDM ÅLDH SOC1 SOC2
UTBM UTBH POL (8.3)

och MODELL 2 utan kommunvariabler, båda med inverterade urvalssannolikheter korrigerade för bortfall som vikter.

A = KV ÅLDM ÅLDH SOC1 SOC2 UTBM UTBH POL (8.4)

De viktigaste resultaten framgår av tabell 8.3.
MODELL 1 ger negativa prediktioner (ned till -0.047) för yngre kvinnor i socialgrupp 3 som endast har obligatorisk utbildning och som vuxit upp i en familj utan politiskt engagemang. Detta gäller oavsett kommuntyp.
För yngre kvinnor i socialgrupp 2 med samma bakgrund blir prediktionerna negativa för medelstora och stora kommuner med medelhög eller hög andel tätortsbefolkning.
Vid test av MODELL 1 mot MODELL 2 visar det sig att kommunvariablerna inte signifikant bidrar till modellens förklaringsgrad.
Även MODELL 2 ger negativa prediktioner för yngre kvinnor i socialgrupp 3 med endast obligatorisk utbildning och utan politiskt engagemang i familjen (-0.029). Prediktionen blir 0.01 för motsvarande kvinnor i socialgrupp 2.
Eftersom den uppskattade effekten för socialgrupp 2 skiljer sig svagt från socialgrupp 3 och de uppskattade effekterna för medelhög och hög utbildning är snarlika har en MODELL 3 prövats. I denna skiljs enbart socialgrupp 1 från socialgrupp 2+3 (SOC=1 flr socialgrupp 1) och utbildning utöver den obligatoriska (UTB=1) från obligatorisk utbildning. MODELL 3 blir således

A = KV ÅLD2 ÅLD3 SOC UTB POL (8.5)

Parameterskattningarna ändras ytterst lite.
Fortfarande erhålles negativa prediktioner för samma delgrupp yngre kvinnor i socialgrupp 2 och 3 som tidigare (-0.021). Socialgrupp 1 är liten (ca 8% i urvalet) varför parameterskattningen för denna grupp blir osäker. Socialgruppsindelningen har därför tagits bort i MODELL 4 trots att SOC-variabeln gav signifikans vid partiellt t-test (se tab. 8.3).

A = KV ÅLD2 ÅLD3 UTB POL (8.6)

Parameterskattningarna ändras obetydligt. Fortfarande kvarstår problemet med negativa prediktioner (-0.024 för yngre kvinnor då UTB=0 och POL=0. Se tabell 8.4).

Tabell 8.3. Modellerna 1, 2 och 3 enligt (8.3), (8.4) och (8.5). (Resultatens osäkerhet underskattas p.gr.a. urvalsformen och svarsbortfallet.)

| | Test av hela modellen p-värde | Determination $R^2$ | Reststandardavvikelse s |
|---|---|---|---|
| MODELL 1 | 0.0001 | 0.0714 | 0.363 |
| MODELL 2 | 0.0001 | 0.0687 | 0.363 |
| MODELL 3 | 0.0001 | 0.067 | 0.363 |

Jämförelse modell 1/modell 2

Test av $H_0$: alla INV- och TÄT-param. = 0
p-värde (formel (6.4)) $\geq 0.10$

| | Modell 1 | Modell 2 | | Modell 3 | |
|---|---|---|---|---|---|
| Variabel | Param. est. | Param. est. | $H_0$: param.=0 p-värde för t-test | Param. est. | $H_0$:param.=0 p-värde för t-test |
| Intercept | 0.099 | 0.040 | | 0.044 | |
| INV2 | -0.032 | | | | |
| INV3 | -0.030 | | | | |
| TÄTM | -0.019 | | | | |
| TÄTH | -0.044 | | | | |
| KV | -0.070 | -0.069 | 0.0001 | -0.065 | 0.0001 |
| ÅLDM | 0.083 | 0.082 | 0.0001 | 0.085 | 0.0001 |
| ÅLDH | 0.120 | 0.120 | 0.0001 | 0.125 | 0.0001 |
| SOC1 | 0.104 | 0.106 | 0.0002 | 0.089 (SOC) | 0.0001 |
| SOC2 | 0.027 | 0.029 | 0.0089 | | |
| UTBM | 0.073 | 0.067 | 0.0007 | | |
| UTBH | 0.097 | 0.089 | 0.0001 | 0.087 (UTB) | 0.0001 |
| POL | 0.123 | 0.125 | 0.0001 | 0.126 | 0.0001 |

Tabell 8.4. Modell 4 enligt (8.6). (Resultatens osäkerhet underskattade p.gr.a. urvalsformen och svarsbortfallet.) Test av hela modellen $p<0.001$, determinationen $R^2=0.061$, reststandardavvikelse $s=0.364$.

| Variabel | Parameter est. | Medelfel | $H_0$:param.=0 p-värde för t-test |
|---|---|---|---|
| Intercept | 0.047 | | |
| KV | -0.071 | 0.015 | 0.0001 |
| ÅLD2 | 0.089 | 0.017 | 0.0001 |
| ÅLD3 | 0.123 | 0.020 | 0.0001 |
| UTB | 0.101 | 0.016 | 0.0001 |
| POL | 0.126 | 0.017 | 0.0001 |

Modellerna 1, 2, 3 och 4 ovan har redovisats utförligt trots att de visat sig vara inadekvata p.gr.a. att de ger negativa prediktioner för vissa delgrupper. Detta har gjorts för att demonstrera att en modell kan vara inadekvat trots att alla koefficienter har "rätt" tecken och är av rimlig storlek, testet av hela modellen har givit stark signifikans och koefficienternas medelfel varit litet vilket gjort alla test av enskilda koefficienter starkt signifikanta.
Om inte modellerna givit negativa prediktioner skulle hela residualmönstret granskats innan man varit mogen för slutbedömningen av modellerna.

## 8.13 Ytterligare modifiering - samspelstermer

De rent additiva modellerna 1-4 kunde inte på ett adekvat sätt beskriva hur aktiviteten genom partier varierar mellan delgrupper. Det är därför dags att pröva om modellen kan förbättras av att man inför interaktionstermer (samspelstermer). Ett samspel mellan exempelvis kön och utbildning innebär att utbildningens effekt på sannolikheten för politisk aktivitet genom partier är lika stor för män och kvinnor. (Samspelsbegreppet förklaras i kap. 5.9.)

Det är önskvärt att modellen innehåller så få samspelstermer som möjligt eftersom den annars blir tung och svåröverskådlig.

En samtidig uppdelning efter socialgrupp (1 och 2+3) och en annan variabel ger ett antal ytterst små delgrupper (t.ex. personer i socialgrupp 1 uppvuxna i familj med politiskt engagemang). Det är därför betydande risk för att enstaka hithörande samspelseffekter av en slump blir mycket starka. En genomförd skattning av en modell med alla individvariabler och alla parvisa samspel bekräftar detta. Socialgruppsindelningen stryks därför. Det visade sig även vid prövning av ett antal olika modeller att samspelet mellan utbildning och kön var svaga och insignifikanta. Jämförelsen gjordes härvid mot MODELL 5 enligt (8.7).

$$A = KV\ \ ÅLDM\ \ ÅLDH\ \ UTB\ \ POL\ \ UTB*ÅLDM\ \ UTB*ÅLDH\ \ POL*ÅLDM\ \ POL*ÅLDH\ \ POL*KV\ \ POL*UTB \qquad (8.7)$$

Denna modell som redovisas i tabell 8.5 har testats mot olika alternativ vilka skiljer sig från (8.7) på olika sätt

- inget samspel vare sig utbildning*ålder eller politik*ålder ($p<0.001$)
- inget samspel utbildning*ålder ($p<0.01$)
- inget samspel politik*ålder ($p<0.001$)
- inget samspel politik*kön ($p<0.001$)

Alla test gav starka nominella signifikanser. Däremot är inte samspelet UTB*POL signifikant på 5%-nivån ($p=0.09$). Detta samspel är emellertid ett av de första man på saklogiska grunder bör ta med varför det behålles i modellen. Tack vare samspelstermerna har de negativa prediktionerna eliminerats. Lägsta predicerade sannolikhet är 0.019 för yngre kvinnor med endast obligatorisk utbildning och utan politiskt engagemang i familjen under uppväxten.

Tecknen på alla termer förefaller logiska. Determination $R^2=0.075$ och reststandardavvikelse $s=0.362$ skiljer sig obetydligt från motsvarande mått för en rad andra modeller som är

snarlika, men dessa mått är som tidigare framhållits föga adekvata för modeller med binär responsvariabel. Viktigt är däremot att genomsnittlig absolut avvikelse mellan predikterad sannolik och observerad relativ frekvens i de olika delgrupperna blir låg. Man erhåller MAD=0.028. Avvikelserna för olika grupper framgår av tabell 8.9.

Tabell 8.5. MODELL 5 enligt (8.7). (Resultatens osäkerhet underskattas p.gr.a. urvalsformen och bortfall.)

Test av hela modellen p=0.0001
Determination $R^2$=0.075
Reststandardavvikelse s=0.362

| Variabel | Param.est. | Medelfel | $H_0$:param.=0 p-värde |
|---|---|---|---|
| Intercept | | | |
| KV | -0.042 | 0.017 | 0.013 |
| ÅLDM | 0.056 | 0.030 | 0.066 |
| ÅLDH | 0.068 | 0.030 | 0.025 |
| UTB | 0.089 | 0.028 | 0.002 |
| UTB*ÅLDM | -0.011 | 0.037 | 0.765 |
| UTB*ÅLDH | 0.101 | 0.041 | 0.013 |
| POL | 0.151 | 0.045 | 0.001 |
| POL*ÅLDM | 0.142 | 0.041 | 0.001 |
| POL*ÅLDH | 0.047 | 0.049 | 0.334 |
| POL*KV | -0.116 | 0.035 | 0.001 |
| POL*UTB | -0.066 | 0.039 | 0.090 |

Tabell 8.6. Variansinflationsfaktorer (VIF) för MODELL 5.

| Term | VIF |
|---|---|
| KV | 1.30 |
| ÅLDM | 4.03 |
| ÅLDH | 3.40 |
| UTB | 4.29 |
| POL | 4.10 |
| UTB*ÅLDM | 4.27 |
| UTB*ÅLDH | 2.56 |
| POL*KV | 1.95 |
| POL*ÅLDM | 2.74 |
| POL*ÅLDH | 2.35 |
| POL*UTB | 1.79 |

Variansinflationsfaktorerna (se kap. 6.13) ges i tabell 8.6 och är inte kritiskt stora för någon parameterskattning. Modellen skulle i och för sig kunna vara en rimlig förklaringsmodell men kan inte göra anspråk på att vara mera än en modell för beskrivning och prediktion.
Referensgruppen med alla indikatorvariabler=0 består av yngre män utan utbildning över folkskola/grundskola och utan erfarenhet av politiskt engagemang i familjen under uppväxttiden. Predikterat värde för denna grupp ges av interceptet 0.061. Värdet för exempelvis medelålders kvinnor med utbildning över folkskola men utan erfarenhet av politiskt engagemang i familjen under uppväxttiden ges av summan av koefficienterna för intercept, KV, ÅLDM, UTB och UTB*ÅLDM vilken blir 0.061-0.042+0.056+0.089-0.011=0.153.
Beskrivningen av modellens resultat ges nedan i termen av "effekter" för UTB och POL trots att det inte är fråga om en förklaringsmodell.

1. Predikterade sannolikheter för personer med enbart folkskola/grundskola och utan politiskt engagemang i familjen under uppväxttiden. Sannolikheterna beräknas ur termerna

KV=-0.042 ÅLDM=0.056
ÅLDH=0.068

och blir

| | Yngre | Medelålders | Äldre |
|---|---|---|---|
| Män | 0.061 | 0.117 | 0.129 |
| Kvinnor | 0.019 | 0.075 | 0.087 |

2. Utbildningseffekter då politiskt engagemang i familjen saknats.

UTB=0.089 UTB*ÅLDM=0.011
UTB*ÅLDH=0.101

Tabellen anger hur mycket predikterade värden höjs för personer med utbildning över folkskola/grundskola.

| Yngre | Medelålders | Äldre |
|---|---|---|
| 0.089 | 0.078 | 0.190 |

3. Effekter av politiskt engagemang i familjen för personer med enbart folkskola/grundskola.

POL=0.151 POL*ÅLDM=0.142
POL*ÅLDH=0.047
POL*KV =-0.116

| | Yngre | Medelålders | Äldre |
|---|---|---|---|
| Män | 0.151 | 0.293 | 0.198 |
| Kvinnor | 0.035 | 0.177 | 0.082 |

4. Kombinerad effekt av utbildning och politiskt engagemang i familjen.

UTB=0.089 UTB*POL=-0.066
POL=0.151 UTB*ÅLDM=-0.011
UTB*ÅLDH=0.101
POL*ÅLDM=0.142
POL*ÅLDH=0.047
POL*KV =-0.116

| | Yngre | Medelålders | Äldre |
|---|---|---|---|
| Män | 0.174 | 0.305 | 0.322 |
| Kvinnor | 0.058 | 0.189 | 0.206 |

De predikterade sannolikheterna för politisk aktivitet genom partier för olika delgrupper framgår av tabell 8.7.

Tabell 8.7. Predikterade sannolikheter för politisk aktivitet genom partier enligt MODELL 5.

| | | | Yngre | Medelålders | Äldre |
|---|---|---|---|---|---|
| UTB=0 | POL=0 | Män | 0.06 | 0.12 | 0.13 |
| | | Kvinnor | 0.02 | 0.08 | 0.09 |
| UTB=1 | POL=0 | Män | 0.15 | 0.20 | 0.32 |
| | | Kvinnor | 0.11 | 0.15 | 0.28 |
| UTB=0 | POL=1 | Män | 0.21 | 0.41 | 0.33 |
| | | Kvinnor | 0.05 | 0.25 | 0.17 |
| UTB=1 | POL=1 | Män | 0.24 | 0.42 | 0.45 |
| | | Kvinnor | 0.08 | 0.26 | 0.29 |

## 8.14 Linjära sannolikhetsmodellen - sammanfattning

Det är svårt att finna en entydig adekvat modell för att beskriva och i ännu högre grad för att förklara hur den politiska aktiviteten genom partier varierar.
En första grov ansats i föregående avsnitt bestod i en modell med både individvariabler och kommunvariabler. Modellen som var linjär i både parametrar och variabler tyder på att kommunvariablerna hade ytterst liten systematisk inverkan på den politiska aktiviteten genom partier. Dessa uteslöts därför i det fortsatta modellbyggnadsarbetet.
Bästa modell, linjär i individvariablerna gav (liksom alla andra prövade sådana modeller) negativa predikterade sannolikheter för vissa kategorier individer.

Modellen kompletterades därför med första ordningens samspelstermer vilka visade sig starkt signifikanta och medförde att de negativa predikterade sannolikheterna försvann. Slutmodellen är lättast att tolka om man tänker sig kön och ålder som klassificeringsvariabler medan utbildning och politiskt engagemang i familjen betraktas om "effektvariabler". För personer med endast obligatorisk utbildning och utan politiskt engagemang i familjen ligger de predikterade sannolikheterna för politisk aktivitet genom partier långt under genomsnittet för populationen. Kvinnorna som tillhör denna kategori ligger under männens aktivitetsnivå och yngsta åldersgruppen (oavsett kön) ligger under de övriga åldersgruppernas nivå.
Utbildning utöver den obligatoriska för personer utan politiskt engagemang i familjen höjer den predikterade sannolikheten för politisk aktivitet och då i särskilt hög grad för den äldsta åldersgruppen.
Politiskt engagemang i familjen för personer med enbart obligatorisk utbildning höjer den predikterade sannolikheten för politisk aktivitet. Effekten är betydande utom för yngre kvinnor. Den är större för män än för kvinnor och oavsett kön lägst i yngsta åldersgruppen (och märkligt nog allra störst bland medelålders personer).
För personer med både politiskt engagemang i familjen och utbildning utöver den obligatoriska är mönstret i stort sett detsamma som för personer med politiskt engagemang i familjen men utan annan utbildning än den obligatoriska. Effekten är dock lika stor för äldre som för medelålders.

## 8.15 Indata till logitprogrammen

Vare sig SAS-programmet LOGIST eller BMDP-programmet PLR ger möjlighet till vägning av observationer vid estimationen. Som indata användes därför en frekvenstabell (tabell 8.7) i vilken vägning med inverterade urvalssannolikheter (justerade för svarsbortfall) redan genomförts. Härvid "standardiserades" vikterna så att cellfrekvenserna summerades till an-

talet personer som givit kompletta svar. Cellfrekvenserna var ej avrundade till heltal utan gavs med 2 decimaler. Som synes av tabell 8.8 skiljer sig vägda och ovägda frekvensen aktiva mycket kraftigt åt inom vissa celler (understrykning).

Tabell 8.8 Vägd frekvenstabell över antalet personer som varit politiskt aktiva genom partier.

K=Kvinna Å=åldersklass U=utbildning P=politiskt engagemang
A=antal aktiva N=antal personer

| KÅUP | Vägt | | | Ovägt | |
|---|---|---|---|---|---|
| | A | N | A/N | A/N | N |
| 0000 | 4.82 | 113.71 | 0.042 | 0.045 | 112 |
| 0001 | 6.80 | 24.21 | 0.281 | 0.304 | 23 |
| 0010 | 39.17 | 264.78 | 0.148 | 0.123 | 228 |
| 0011 | 14.72 | 92.99 | 0.158 | 0.200 | 85 |
| 0100 | 21.51 | 177.78 | 0.121 | 0.160 | 213 |
| 0101 | 14.97 | 38.43 | 0.390 | 0.415 | 41 |
| 0110 | 36.81 | 195.12 | 0.189 | 0.199 | 166 |
| 0111 | 54.20 | 114.70 | 0.473 | 0.307 | 75 |
| 0200 | 29.01 | 204.66 | 0.142 | 0.171 | 251 |
| 0201 | 17.76 | 54.65 | 0.325 | 0.406 | 69 |
| 0210 | 35.48 | 110.40 | 0.321 | 0.338 | 68 |
| 0211 | 12.30 | 26.84 | 0.458 | 0.639 | 36 |
| 1000 | 1.96 | 89.30 | 0.022 | 0.019 | 106 |
| 1001 | 0.65 | 10.78 | 0.060 | 0.133 | 15 |
| 1010 | 25.69 | 219.66 | 0.117 | 0.090 | 201 |
| 1011 | 10.80 | 73.41 | 0.147 | 0.173 | 75 |
| 1100 | 10.33 | 145.46 | 0.071 | 0.149 | 167 |
| 1101 | 7.26 | 26.85 | 0.270 | 0.225 | 40 |
| 1110 | 22.19 | 139.12 | 0.160 | 0.127 | 126 |
| 1111 | 10.07 | 60.19 | 0.167 | 0.188 | 48 |
| 1200 | 11.74 | 150.15 | 0.078 | 0.105 | 191 |
| 1201 | 5.07 | 35.53 | 0.143 | 0.189 | 53 |
| 1210 | 15.89 | 64.30 | 0.247 | 0.279 | 43 |
| 1211 | 7.70 | 21.91 | 0.351 | 0.250 | 20 |

## 8.16 Logitprogrammet PLR

BMDP-programmet PLR utnyttjar maximumlikelihoodmetoden vid skattning av logitmodellen. Programmet förutsätter obundet slumpmässigt urval av individer varför medelfel och testkarakteristikor inte heller för detta program är adekvata för ett tvåstegsurval.

## 8.17 Ny parametrisering

I PLR genereras en annan kodning av indikatorvariabler (kallade designvariabler i BMDP-manualen) än den som tidigare använts i denna rapport.

$$KV = \begin{cases} -1 & \text{män} \\ 1 & \text{kvinnor} \end{cases}$$

$$BTU = \begin{cases} -1 & \text{folkskola/grundskola} \\ 1 & \text{utbildning över folkskola/grundskola} \end{cases}$$

$$POL = \begin{cases} -1 & \text{inget politiskt engagemang i familjen} \\ 1 & \text{politiskt engagemang i familjen} \end{cases}$$

$$ÅLD1 = \begin{cases} -1 & \text{yngre} \\ 0 & \text{medelålders} \\ 1 & \text{äldre} \end{cases}$$

$$ÅLD2 = \begin{cases} -1 & \text{yngre} \\ 1 & \text{medelålders} \\ 0 & \text{äldre} \end{cases}$$

$\beta$-parametrarna i modellen

$$\ln(P/1-P) = \alpha+\beta_k \cdot KV+\beta_{Å1} \cdot ÅLD1+\beta_{Å2} \cdot ÅLD2+\beta_U \cdot UTB+\beta_p \cdot POL \qquad (8.8)$$

förekommer nu på grund av designvariablernas kodning i avvikelseform. Summan (och medelvärdet) av en viss effekt över de olika klasserna för variabeln blir 0.
Om könseffekten för kvinnor är exempelvis -0.11 blir den 0.11 för män.

Parametrarna i logitmodellen är oavsett parametrisering inte jämförbar med den vanliga linjära regressionsmodellens parametrar. Det finns därför ingen anledning att räkna om logitparametrarna till en annan parametrisering.

## 8.18 Prövade logitmodeller

Inledningsvis skattades MODELL 6 enligt (8.8) symboliskt skriven

$$\ln P/(1-P) = KV \quad ÅLD \quad UTB \quad POL$$

Hur väl modellen beskriver datastrukturen anges av ett likelihood-ratio $\chi^2$-test. Detta gav $\chi^2=34.75$ med df=18 svarande mot ett p-värde på 0.010. Ett lågt $\chi^2$-värde med åtföljande högt p-värde är tecken på god anpassning. MODELL 6 är således föga adekvat även med tanke på att signifikanserna överdrivs på grund av att programmet förutsätter obundet slumpmässigt urval. Det kan ändå vara av intresse att se hur väl MODELL 6 predikterar sannolikheterna för politisk aktivitet jämfört med MODELL 5 d.v.s. bästa linjära regressionsmodell med samspelstermer. Här ger logitmodellen betydligt sämre resultat (tabell 8.9). Parameterskattningarna återfinnes i tabell 8.10.

Olika logitmodeller med interaktionstermer har prövats stegvis.

MODELL

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 6. | logit = | KV | ÅLD | UTB | POL | | | |
| 7. | logit = | KV | ÅLD | UTB | POL | UTB*POL | | |
| 8. | logit = | KV | ÅLD | UTB | POL | UTB*POL | KV*POL | |
| 9. | logit = | KV | ÅLD | UTB | POL | UTB*POL | KV*POL | ÅLD*POL |
| 10. | logit = | KV | ÅLD | UTB | POL | UTB*POL | KV*POL | ÅLD*POL |
| | | | ÅLD*UTB | | | | | |

Tabell 8.9. Prediktion av politisk aktivitet genom partier med olika modeller.

MODELL 5: Linjär regressionsmodell med interaktionstermer.
MODELL 6: Logitmodell med huvudeffekter.
MODELL 9: Logitmodell med interaktionstermer.
MODELL 10: Logitmodell med interaktionstermer.

| Grupp KÅUP | Obs frekv. (vägd) | Mod.5 Pred. | Mod.5 Resid. | Mod.6 Pred. | Mod.6 Resid. | Mod.9 Pred. | Mod.9 Resid. | Mod.10 Pred. | Mod.10 Resid. |
|---|---|---|---|---|---|---|---|---|---|
| 1000 | 0.022 | 0.019 | 0.003 | 0.039 | -0.017 | 0.041 | -0.019 | 0.038 | -0.016 |
| 0000 | 0.042 | 0.061 | -0.019 | 0.066 | -0.024 | 0.059 | -0.017 | 0.054 | -0.012 |
| 1001 | 0.060 | 0.054 | 0.006 | 0.084 | -0.024 | 0.084 | -0.024 | 0.074 | -0.014 |
| 1100 | 0.071 | 0.075 | -0.004 | 0.075 | -0.004 | 0.065 | 0.006 | 0.078 | -0.007 |
| 1200 | 0.078 | 0.087 | -0.009 | 0.097 | -0.019 | 0.099 | -0.021 | 0.089 | -0.011 |
| 1010 | 0.117 | 0.108 | 0.009 | 0.082 | 0.035 | 0.104 | 0.013 | 0.106 | 0.011 |
| 0100 | 0.121 | 0.117 | 0.004 | 0.124 | -0.003 | 0.092 | 0.029 | 0.111 | 0.010 |
| 1201 | 0.143 | 0.169 | -0.026 | 0.195 | -0.052 | 0.189 | -0.046 | 0.170 | -0.027 |
| 0200 | 0.142 | 0.129 | 0.013 | 0.157 | -0.015 | 0.139 | 0.003 | 0.125 | 0.017 |
| 1011 | 0.147 | 0.077 | 0.070 | 0.167 | -0.020 | 0.104 | 0.043 | 0.105 | 0.042 |
| 0010 | 0.148 | 0.150 | -0.002 | 0.134 | 0.016 | 0.146 | 0.002 | 0.148 | 0.000 |
| 0011 | 0.158 | 0.235 | -0.077 | 0.258 | -0.100 | 0.216 | -0.058 | 0.221 | -0.063 |
| 1110 | 0.160 | 0.153 | 0.007 | 0.150 | 0.010 | 0.160 | 0.000 | 0.147 | 0.013 |
| 1111 | 0.167 | 0.264 | -0.103 | 0.286 | -0.117 | 0.251 | -0.082 | 0.237 | -0.069 |
| 0110 | 0.189 | 0.195 | -0.006 | 0.235 | -0.046 | 0.219 | -0.030 | 0.202 | -0.013 |
| 1210 | 0.247 | 0.277 | -0.030 | 0.190 | 0.057 | 0.232 | 0.015 | 0.255 | -0.008 |
| 1101 | 0.270 | 0.252 | 0.018 | 0.155 | 0.115 | 0.210 | 0.060 | 0.239 | 0.031 |
| 0001 | 0.281 | 0.212 | 0.069 | 0.138 | 0.143 | 0.179 | 0.102 | 0.162 | 0.119 |
| 0210 | 0.321 | 0.319 | 0.002 | 0.289 | 0.032 | 0.308 | 0.013 | 0.335 | -0.014 |
| 0201 | 0.325 | 0.327 | -0.002 | 0.296 | 0.029 | 0.357 | -0.032 | 0.332 | -0.007 |
| 1211 | 0.351 | 0.293 | 0.058 | 0.346 | 0.005 | 0.278 | 0.073 | 0.261 | 0.090 |
| 0101 | 0.390 | 0.410 | -0.020 | 0.242 | 0.148 | 0.387 | 0.003 | 0.432 | 0.034 |
| 0211 | 0.458 | 0.451 | 0.007 | 0.479 | -0.021 | 0.412 | 0.046 | 0.460 | -0.002 |
| 0111 | 0.473 | 0.422 | 0.051 | 0.410 | 0.063 | 0.444 | 0.029 | 0.429 | 0.044 |
| MAD* | | | 0.026 | | 0.046 | | 0.032 | | 0.028 |

*MAD = Mean absolute deviation = $\frac{1}{n}\Sigma|\text{Obs-Pred}|$

Testproceduren gav följande resultat:

| Modell | Term tillagd | Förbättring $\chi^2$ | p-värde | Modellanpassning $\chi^2$ | p-värde |
|---|---|---|---|---|---|
| 6 | (referensmodell) | | | 34.75 | 0.010 |
| 7 | UTB*POL | 7.93 | 0.005 | 26.83 | 0.061 |
| 8 | KV*POL | 3.84 | 0.050 | 22.99 | 0.114 |
| 9 | ÅLD*POL | 5.75 | 0.056 | 17.24 | 0.243 |
| 10 | ÅLD*UTB | 4.11 | 0.105 | 13.14 | 0.359 |

MODELLERNA 9 och 10 har undersökts närmare. Parameterskattningarna återfinns i tabell 8.10. Av dessa parametrar kan man dra slutsatser om olika variablers effekt på logit (P) d.v.s. på log [P/(1-P)]. Parametervärdena i sig är skäligen ointressanta. Till och med för den enkla MODELL 6 utan samspelstermer kan man direkt av parameterskattningarna utläsa ytterligt lite om effekten på sannolikheten P av olika variabler. Som framhållits i kap. 7.3 samspelar variablerna i effekten på sannolikheten P trots att variablerna har en additiv effekt på logit (P). Som illustration visas i tabell 8.11 utbildningseffekten på sannolikheten P för män och kvinnor i skilda åldersgrupper som ej vuxit upp i familj med politiskt engagemang.

Dessa skattningar erhålles ej från PLR-programmet utan måste räknas fram ur tabell 8.9. Utbildningseffekten för yngre kvinnor erhålles som skillnaden mellan predikterade P-värden för gruppen 1010 (med utbildning över grundskola/folkskola) och gruppen 1000 (utan utbildning) d.v.s. som 0.082-0.039= =0.043.

Utbildningseffekten är som synes både köns- och åldersberoende med störst effekt för äldsta åldersgruppen.

Utbildningseffekten på P enligt MODELL 10 för samma kategorier, framgår också av tabell 8.11 och är beräknade från tabell 8.9 på precis samma sätt som ovan. Observera att ett teckenstudium av koefficienterna för logitmodellen 10 absolut inte säger någonting om variablernas samspelseffekter på sannolikheten P. Jämförelser av parametertecken i regressionsmodellen 6 och logitmodellen 10 (som innehåller exakt samma regressorer) är således utan mening.

Tabell 8.10. Parameterskattningar för logitmodellerna 6, 9 och 10. (Resultatens osäkerhet underskattas p.gr.a. urvalsform och svarsbortfall.)

| Variabel | MODELL 6 | | MODELL 9 | | MODELL 10 | |
|---|---|---|---|---|---|---|
| | Param. est. | Medel-fel | Param. est. | Medel-fel | Param. est. | Medel-fel |
| KV | -0.276 | 0.059 | -0.313 | 0.063 | -0.317 | 0.064 |
| ÅLD (0)[a] | -0.554 | 0.086 | -0.574 | 0.094 | -0.626 | 0.116 |
| (1) | 0.417 | 0.085 | 0.370 | 0.095 | 0.378 | 0.098 |
| (2) | 0.137 | 0.076 | 0.204 | 0.081 | 0.248 | 0.091 |
| UTB | 0.391 | 0.062 | 0.312 | 0.067 | 0.333 | 0.072 |
| UTB*ÅLD (0)[a] | | | | | 0.041 | 0.110 |
| (1) | | | | | 0.118 | 0.090 |
| (2) | | | | | -0.159 | 0.087 |
| POL | 0.408 | 0.059 | 0.404 | 0.069 | 0.407 | 0.069 |
| POL*ÅLD (0)[a] | | | -0.091 | 0.108 | -0.105 | 0.094 |
| (1) | | | -0.102 | 0.095 | -0.088 | 0.095 |
| (2) | | | 0.193 | 0.081 | 0.193 | 0.080 |
| POL*KV | | | -0.121 | 0.063 | -0.125 | 0.064 |
| POL*UTB | | | -0.194 | 0.067 | -0.180 | 0.067 |
| Constant | | | -1.578 | | -1.569 | |

a) Denna parameter för yngsta åldersklassen svarar ej mot någon designvariabel och ges ej i datautskriften. Beräknas som "parametersumman för medelålders (1) och äldre (2) med omvänt tecken".

Tabell 8.11. Utbildningseffekter då POL=0 enligt logitmodellerna 6 och 10.

MODELL 6

| | Yngre | Medelålders | Äldre |
|---|---|---|---|
| Män | 0.068 | 0.111 | 0.132 |
| Kvinnor | 0.043 | 0.075 | 0.093 |

MODELL 10

| | Yngre | Medelålders | Äldre |
|---|---|---|---|
| Män | 0.094 | 0.091 | 0.210 |
| Kvinnor | 0.068 | 0.069 | 0.166 |

Val mellan prövade modeller. Tabell 8.9 visar att logitmodellerna ger större genomsnittligt absolut prediktionsfel (MAD) än den linjära regressionsmodellen 5 med samspelseffekter. Eftersom man dessutom, i motsats till logitmodellen, enkelt kan tolka parametrarna i en linjär regressionsmodell är det självklart att denna väljs som slutgiltig modell.

I ett senare kapitel prövas en produktmodell för att beskriva den politiska aktivitetens variationer mellan celler i en flervägs kontingenstabell.

# 9. KONTINGENSTABELLER

## 9.1 Modelltyper

Teorin för analys av samband och samvariation mellan kategoriska variabler, vanligen redovisade i en kontingenstabell, har länge varit mager jämfört med teorin för analys av kontinuerliga variabler. Den förra har i huvudsak varit begränsad till olika associationsmått och till $\chi^2$-test av oberoende i en kontingenstabell. Under det senaste 15-talet år har det försiggått en synnerligen omfattande utveckling av teori för modellbaserad analys av samband mellan kategoriska variabler. De allra senaste åren har denna teori blivit mera allmänt tillgänglig i handböcker. Numera finns också enstaka datorprogram i standardprogrampaketen för analys av sådana modeller. Det förekommer två huvudtyper av modeller: asymmetriska och symmetriska. De förra har likheter med regressionsmodeller, de senare med korrelationsmodeller.
Analysen av modeller för kategoriska data har utgått från olika estimationsmetoder: "maximum likelihood"(ML), generaliserad minsta kvadrat (GLS efter engelska termen) och "minimum discrimination information". Här kommer GLS- och ML-metoder att beröras.

## 9.2 Asymmetriska modeller

GLS-ansatsen. Den asymmetriska modelltypen har en beroende variabel och ett antal förklaringsvariabler eller faktorer. Grizzle, Starmer and Koch [10] har utvecklat GLS-ansatsen för kategoriska data. De behandlar en vid klass av modeller i vilka väntevärdet för den beroende variabeln anges som en linjär funktion av parametrar för förklaringsvariablerna. Då responserna är binära kan den beroende variabeln utgöras av proportionen "positiva" responser (p) i cellerna i kontingenstabellen eller av en funktion av proportionen "positiva". De i kapitel 7 och 8 behandlade logitmodellerna i vilka den beroende variabeln definieras som $\ln[p/(1-P)]$ inryms i denna klass.
Grizzle, Starmer and Koch har beskrivit hur GLS-metoden används för att anpassa den linjära modellen, för att testa hur bra anpassningen är och för att testa hypoteser om parametrarna i modellen.

Metoden kallas ofta GSK-metoden efter författarna. Andra vanliga benämningar vid referens till denna metodik är GLS-ansatsen (generalized least squares) och något oegentligt WLS-ansatsen (Weighted least squares).
Vid användning av asymmetriska modeller är intresset koncentrerat till estimation av parametrar.

Regressionsmodeller. Denna modelltyp är utvecklad för kontinuerlig responsvariabel. Som framgått av tidigare kapitel är den i vissa teoretiska avseenden inadekvat för analys av kategoriska responsvariabler. Använd med omdöme kan den dock, som gramgår av kapitel 8, ge en beskrivning och belysning av sakförhållandena. GSK-ansatsen saknar emellertid de teoretiska brister som vidlåder regressionsmodellerna. Vid estimationen har man utgått från de speciella egenskaperna hos kategoriska variabler (beträffande bl.a. variansegenskaper) och använt estimationsmetoder som garanterat vissa goda estimatorsegenskaper.
Fördelen med regressionsmodellerna ligger i att de är enkla och välkända samt att lätthanterliga standardprogram är allmänt förekommande.
Fördelen med de speciella nyare modelltyperna ligger i att de är teoretiskt adekvata, möjliggör en smidigare och mer långtgående analys av sambandsstrukturer samt att de i termer av parameterskattningar ger något bättre resultat på grund av lämpligare skattningsmetoder.

## 9.3 Symmetriska modeller

Multiplikativ modell. Den andra huvudansatsen behandlar alla variabler symmetriskt som faktorer. Det finns alltså formellt inte någon variabel som utgör responsvariabel. Förväntat antal observationer i en cell i en flerdimensionell kontingenstabell anges som en produkt av parametrar för faktorerna. Det är alltså fråga om en multiplikativ modell. På grund av att man vid skattningen logaritmerar modellerna för att uppnå linearitet talar man emellertid ofta om "loglinjära modeller". Detta tycks ha fått många samhällsvetare att tro att det är fråga om en ganska speciell modelltyp vilket inte är fallet. Vid användning av denna modelltyp är intresset koncentrerat till förekomsten av interaktionseffekter av olika ordning och till graden av

anpassning med olika uppsättning termer. Parametervärden är svårtolkade och av underordnad betydelse.
Då sakproblemet gör det meningsfullt är det möjligt att tolka dessa modeller i termer av en responsvariabel och dess beroende av faktorerna, men i allmänhet utan att parameterskattningar kommer i förgrunden.
Om parametervärden är ett primärt intresse bör man i stället använda de i föregående delavsnitt berörda asymmetriska modellerna.

## 9.4 Litteratur

Den nya metodiken för analys av kontingenstabeller sprids snabbt och det kommer ganska säkert inom några år att krävas att varje någorlunda kvalificerad analytiker av surveydata känner till och förstår dessa nya metoder på samma sätt som det sedan länge krävts att han/hon behärskar och förstår regressionsanalysen.
Det är ont om handböcker om de additiva modellerna (GSK-metoden). En sådan bok har utkommit under 1981 (för fullständiga referenser se referenslistan).

[ 7] Forthofer and Lehnen: Public Program Analysis A New Categorical Data Approach. 1981.
Medelsvår. Utnyttjar matrisalgebra. Analyserar ett antal samhällsvetenskapliga datamaterial.

Tillgången på handböcker är bättre vad beträffar produktmodellen (loglinjära modellen). Ett urval följer:

[ 1] Bishop, Fienberg and Holland: Discrete Multivariate Analysis. Theory and Practice. 1975.
Svår. Betraktas som standardverk inom området. Visar en rad olika modelltyper för speciella problemtyper av synnerligen varierande art. Värdefull som uppslagsverk för empirikern som väl behärskar grunderna.

[ 5] Everitt: The Analysis of Contingency Tables. 1977.
Lätt.

[ 6] Fienberg: The Analysis of Cross-Classified Data. 1977.
Lätt.

[19] Knoke and Burke: Log-Linear Models. 1980.
Lättaste boken. Ger en enkel och klar bild av begreppsapparaten. Rekommenderas ej som stöd

vid användning av modellerna. Sätter ej modellutformning i relation till design för datainsamling vilket bäddar för missbruk.

[24] Payne: The Log-Linear Model for Contingency Tables. 1977.

Medelsvår uppsats i samlingsvolymen "Analysis of Survey Data" Vol. 2.

[27] Upton: The Analysis of Cross-tabulated Data. 1978.

Lätt. Utförlig beskrivning av modellbyggnadsprocessen med hjälp av brittiska valdata. Den bok bland här uppräknade som enligt min mening utgör det bästa stödet för samhällsvetare vid praktiskt modellarbete. Lämplig grund för den mera drivne analytiker som önskar tillgodogöra sig de praktiskt inriktade delarna i Bishop et. al och deras beskrivning av mera speciella modeller.

# 10. LINJÄR MODELL

## 10.1 Allmän form

Modelltypen förutsätter att man har en kategorisk responsvariabel och en eller flera kategoriska förklaringsvariabler eller faktorer. (En kategorisk variabel är genuint kvalitativ eller har definierats för att identifiera klasserna vid klassindelning av en kvantitativ variabel.) Utgångspunkten för analysen utgöres av en kontingenstabell med $k \geq 1$ faktorer. För varje cell (eller population enligt terminologi i SAS-programmet FUNCAT) som definieras vid korsklassificering efter faktorerna observeras antalet observationer i varje responskategori (klass) för responsvariabeln.

Antag tills vidare att responsvariabeln är binär. Värdet 1 anger förekomst av viss egenskap och värdet 0 avsaknad av egenskapen ifråga. Den storhet man intresserar sig för är andelen individer i cellen som har egenskapen ifråga (A=1). Den beroende variabeln utgöres av denna andel (p) i sig eller av en funktion av p såsom logaritmiska oddset (logit) för p definierat av $\log[p/(1-p)]$.

Väntevärdet för denna beroende variabel antas sedan vara linjärt i vissa parametrar som beror av faktorerna (förklaringsvariablerna). Till det yttre ser väntevärdesuttrycket ut som en vanlig regressionsmodell med enbart dummyvariabler. Då den beroende variabeln utgöres av den observerade andelen p i cellen är modellen nära släkt med de linjära regressionsmodeller med interaktionstermer som använts i kapitel 8. Den skiljer sig från dessa genom att modellen avser den observerade andelen p i cellen och inte enskilda individvärden. Den skiljer sig också från standardmodellen för regression genom att man har en för kategoriska data korrekt specifikation av varians och kovarians för den beroende variabeln.

Den beroende variabeln kan utgöras av en linjär, logaritmisk eller exponentiell funktion av responsvariabeln eller av en sammansatt funktion av dessa.

Då responsvariabeln har fler än två nivåer vill man ofta bilda en komplex funktion. Således tillåter GLS-metodiken att man studerar exempelvis rangkorrelationskoefficienter som beroende variabler. Se Forthofer and Lehnen [ 7].

I detta kapitel behandlas endast det enkla fall då den beroende variabeln utgöres av de observerade andelarna inom cellerna. Framställningen blir kortfattad för att ge utrymme för en utförligare behandling av produktmodellen (loglinjära modellen) i följande kapitel.

## 10.2 Linjär modell för p

Låt som exempel faktorn A (kön) ha 2 klasser och faktorn B (utbildning) ha 3 klasser. För varje kombination av A- och B-klasser göres ett antal observationer på en binär responsvariabel H som anger om en viss händelse inträffar eller ej. $p_{ij}$ anger andelen oberoende "försök" av $n_{ij}$ i cell (i,j) som leder till att händelsen inträffar. En modell för $p_{ij}$ av variansanalystyp och utan interaktionstermer kan skrivas

$$p_{ij} = p_{ij}+\varepsilon_{ij} = \mu+a_i+b_j+\varepsilon_{ij} \qquad (10.1)$$

$$E(\varepsilon_{ij})=0$$

$$V(p_{ij}|n_{ij})= p_{ij}(1-p_{ij})/n_{ij}$$

för i=1,2 och j=1,2,3. Här är en genomsnittsnivå medan $a_i$ och $b_j$ anger avvikelser från denna nivå. Medelavvikelsen (ovägd) för A-nivåerna skall vara 0 vilket medför att $a_1+a_2=0$, det vill säga $a_2=-a_1$ . Det räcker därför att definiera en parameter a för nivå $A_1$ vilket medför att huvudeffekten för nivå $A_2$ blir -a .
På motsvarande sätt kan de 3 B-effekterna uttryckas med hjälp av två parametrar: $b_1$ för nivå $B_1$ och $b_2$ för nivå $B_2$ . Effekten för nivå 3 blir då $-b_1-b_2$ .
Med hjälp av designvariabler (indikatorvariabler) kan modellen skrivas på regressionsform. I de tidigare kapitlen har indikatorvariablerna antagit värdena 0 och 1. Eftersom modellen här ges i avvikelseform blir det fråga om värdena 1, 0 och -1 .

A-klassificeringen med 2 klasser anges av designvariabeln (indikatorn)

$$X_2 = \begin{cases} 1 \text{ för klass A1} \\ -1 \text{ för klass A2} \end{cases}$$

(Skälet till att börja variabelnumreringen på 2 kommer strax att framgå.) B-klassificeringen med 3 klasser kräver 2 disignvariabler.

$$X_3 = \begin{cases} 1 & \text{för klass B1} \\ 0 & \text{för klass B2} \\ -1 & \text{för klass B3} \end{cases} \qquad X_4 = \begin{cases} 0 & \text{för klass B1} \\ 1 & \text{för klass B2} \\ -1 & \text{för klass B3} \end{cases}$$

Modellen (10.1) kan nu skrivas på följande sätt för cell $(A_i,B_j)$

$$p_{ij} = \mu+aX_2+b_1X_3+b_2X_4+\varepsilon \qquad (10.2)$$

För exempelvis cellen $(A_1,B_2)$ erhålles $p_{12}=\mu+a_1+b_2+\varepsilon$ .

Skrivsättet kan göras ännu enhetligare om vi inför en designvariabel $X_1$ som alltid antar värdet 1 (oavsett A- och B-nivå) och multiplicerar $\mu$ med denna.

Alltså $p_{ij} = \mu X_1+aX_2+b_1X_3+b_2X_4+\varepsilon$ .

Floran av bokstäder kan skäras ned om man använder ett indicerat $\beta$ för alla parametrar. Låt $\mu=\beta_1$, $a=\beta_2$, $b_1=\beta_3$ och $b_2=\beta_4$ . Då blir modellen

$$p_{ij} = \sum_{1}^{4}\beta_iX_i+\varepsilon \qquad (10.3)$$

Övre summationsgränsen som här är 4 anger antal parametrar i modellen. Symboliskt skulle modellen i SAS-programmet FUNCAT skrivas

MODEL H = A B; (10.4)

Designmatrisen. Numrera nu i stället cellerna i kontingenstabellen löpande enligt nedan.

| | $B_1$ | $B_2$ | $B_3$ |
|---|---|---|---|
| $A_1$ | 1 | 2 | 3 |
| $A_2$ | 4 | 5 | 6 |

Designvariablernas värde (enligt definitioner ovan) kan anges för varje cell i en designmatris. (Tabell 10.1) Dataprogramutskrifter kräver att man kan läsa sådana designmatriser.

Tabell 10.1. Designmatris

| | Designvariabler | | | |
|---|---|---|---|---|
| Cell | Konstant $X_1$ | A-klassif. $X_2$ | B-klassif. $X_3$ | $X_4$ |
| 1 | 1 | 1 | 1 | 0 |
| 2 | 1 | 1 | 0 | 1 |
| 3 | 1 | 1 | -1 | -1 |
| 4 | 1 | -1 | 1 | 0 |
| 5 | 1 | -1 | 0 | 1 |
| 6 | 1 | -1 | -1 | -1 |

Här kan man direkt avläsa från exempelvis rad 3 att väntevärdet för p i cell 3 $(A_1,B_3)$ uttrycks på följande sätt i parametrarna:

$$\beta_1 \cdot 1+\beta_2 \cdot 1+\beta_3(-1)+\beta_4(-1) = \beta_1+\beta_2-\beta_3-\beta_4$$

Interaktionstermer. Komplettera modellen ovan med interaktionstermer.

$$p_{ij} = \mu+a_i+b_j+c_{ij}+\varepsilon_{ij} \qquad (10.5)$$

$c_{ij}$ anger hur mycket $E(p_{ij})$ avviker från vad som förväntas med hänsyn till huvudeffekter av A och B. Denna förväntan är $\mu+a_i+b_j$ . Medelvärdet (och summan) av avvikelserna inom varje rad är 0. Samma sak gäller avvikelserna inom varje kolumn. Således kan alla 6 $c_{ij}$-termerna beräknas om man känner 2 lämpligt valda bland dem. Konventionen består i att uttrycka övriga samspelstermer med hjälp av $c_{11}$ och $c_{12}$. Mönstret för samspelstermerna blir därför som i tabell 10.2.

Tabell 10.2. Samspelstermer.

| | $B_1$ | $B_2$ | $B_3$ |
|---|---|---|---|
| $A_1$ | $c_{11}$ | $c_{12}$ | $-c_{11}$ $-c_{12}$ |
| $A_2$ | $-c_{11}$ | $-c_{12}$ | $c_{11}$ $+c_{12}$ |

Med 2 A-klasser och 2 B-klasser kan alla $c_{ij}$ uttryckas i den enda termen $c_{11}$ enligt nedan:

| | $B_1$ | $B_2$ |
|---|---|---|
| $A_1$ | $c_{11}$ | $-c_{11}$ |
| $A_2$ | $-c_{11}$ | $c_{11}$ |

Designmatrisen vid interaktion. Designmatrisen i tabell 10.1 kan byggas ut till en designmatris för modell (10.5) med interaktion. För varje cell i kontingenstabellen måste det anges om parametern $c_{11}$ skall multipliceras med 1, 0 eller -1 innan den adderas i väntevärdesberäkningen. Samma sak gäller parametern $c_{12}$. (Se tablån över $c_{ij}$-termen ovan.) Designvariabeln $X_2$ bestämmer som nämnts A-klass medan $X_3$ och $X_4$ bestämmer B-klass.

Bilda produkterna $X_2X_3$ och $X_2X_4$ enligt tabell 10.3.

Tabell 10.3. Produkt av designvariabler.

| | | $B_1$<br>$X_3=1$<br>$X_4=0$ | $B_2$<br>$X_3=0$<br>$X_4=1$ | $B_3$<br>$X_3=-1$<br>$X_4=-1$ |
|---|---|---|---|---|
| $A_1$ | $X_2=1$ | $X_2X_3=1$<br>$X_2X_4=0$ | $X_2X_3=0$<br>$X_2X_4=1$ | $X_2X_3=-1$<br>$X_2X_4=-1$ |
| $A_2$ | $X_2=-1$ | $X_2X_3=-1$<br>$X_2X_4=0$ | $X_2X_3=0$<br>$X_2X_4=-1$ | $X_2X_3=1$<br>$X_2X_4=1$ |

En jämförelse med tabell 10.2 visar att produkten $X_2X_3$ i varje cell ger konstanten framför parametern $c_{11}$ medan $X_2X_4$ ger konstanten framför $c_{12}$.
Produkten $X_2X_3$ är i sig en designvariabel. Kalla denna för $X_5$ och kalla produkten $X_2X_4$ för $X_6$.

Designmatrisen för modell (10.5) med interaktion kan nu skrivas enligt tabell 10.4. Modellen för $p_{ij}$ skrivs nu:

$$p = \sum_{1}^{6} \beta_i X_i \qquad (10.6)$$

och i programmet FUNCAT

MODEL H = A B A*B; (10.7)

Observera att ordningsföljden mellan disignvariablerna bestäms av ordningsföljden mellan termerna i MODEL-satsen.

Tabell 10.4. Designmatris för modell (10.5) = modell (10.6).

| | Designvariabler | | | | | |
|---|---|---|---|---|---|---|
| | Konstant | A-klass | B-klass | | Interaktion | |
| Cell | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ (=$X_2X_3$) | $X_6$ (=$X_2X_4$) |
| 1 | 1 | 1 | 1 | 0 | 1 | 0 |
| 2 | 1 | 1 | 0 | 1 | 0 | 1 |
| 3 | 1 | 1 | -1 | -1 | -1 | -1 |
| 4 | 1 | -1 | 1 | 0 | -1 | 0 |
| 5 | 1 | -1 | 0 | 1 | 0 | -1 |
| 6 | 1 | -1 | -1 | -1 | 1 | 1 |

Av exempelvis rad 5 kan utläsas att väntevärdet för p blir $\beta_1-\beta_2+\beta_4-\beta_6$ eller i ursprungliga parameterbeteckningar $\mu-a+b_2-c_{12}$ .

Designmatrisen används vid tolkning av parameterskattningarna i en datautskrift vilket vi skall se prov på nedan. Den används också vid estimationen av modellen vilket emellertid inte skall behandlas här.

## 10.3 Politisk aktivitet

För att analysera den politiska aktiviteten genom partier ansättes motsvarigheten till regressionsmodellen 5 med interaktionseffekter. Symboliskt skrivs denna (MODELL 11).

A = KV ÅLD UTB POL UTB*ÅLD POL*KV POL*ÅLD UTB*POL (10.7)

på samma sätt som för modell 5.
Modellen har skattats med generaliserad minstakvadratmetod vilken asymptotiskt är ekvivalent med maximumlikelihoodmetoden. GLS-skattningar är därför asymptotiskt effektiva. Vid skattningen har SAS-programmet FUNCAT använts.

## 10.4 FUNCAT-programmet

FUNCAT är för närvarande (1982) det enda programmet i de vanligare standardprogrampaketen som är utformat för analys av kategoriska data med hjälp av linjära modeller. Härvid bortses från logitprogrammen som enbart tillåter ln[P/(1-P)] som beroende variabel medan FUNCAT ger en stor valfrihet i detta avseende.
FUNCAT kräver djupare kunskaper av användaren än ett vanligt regressionsprogram gör. I SAS-manualen förutsätts att läsaren kan tillräckligt mycket statistisk teori för att i matematiskt symbolspråk kunna specificera vilken funktion av andelen p som skall utgöra beroende variabel, samt för att kunna tolka parameterskattningarna.
Dessa problem utreds enklast i anslutning till program och programutskrift för analys av MODELL 11. FUNCAT-proceduren fordrar en kontingenstabell som indata vilken i detta fall bildas från rådata med samtidig vägning för urvalssannolikheter och svarsbortfall. FUNCAT ger automatiskt parametrarna i avvikelseform. Programmet kan se ut som följer.

```
Rad
 10 /JOB
 20 // EXEC SAS
 30 //IN DD osv....
 40 //SYSIN DD *
 50 DATA GRUND;
 60   INFILE IN;
 70   INPUT V367 17 V368 18 V520 19-20
 80      V522 21 V544 23-27 A 28;
 90   POL=5;
100   IF V367=1 THEN POL=1;
```

Ja och nej har här kodats 1 resp. 5 i stället för 1 resp. 0 som tidigare. Detta för att designvariabeln för POL som antar värdet 1 för lägsta POL-värdet och -1 för det högsta skall indikera ja-svar med värdet 1. Eljest erhålles omvänt tecken för

POL-parametern i modell 11 jämfört med motsvarande parameter i modell 5. (Detta klarnar då vi återkommer till designmatrisen nedan.)

```
110    KV=5;
120    IF V522=1 THEN KV=1;
130    UTB=5;
140    IF V368=2 OR V368=3 THEN UTB=1;
150    VIKT=V544*2452/263098;
```

Variabeln V544 anger inverterade inklusionssannolikheten inom kommunen korrigerad för individbortfall. Vissa svarande har dock partiellt svarsbortfall på vissa variabler varför endast 2452 observationer (med variabelsumman 263098) som är fullständiga kan användas. V544 tar ej hänsyn till urvalsfraktionen av kommuner i första urvalssteget. Variabeln V544 måste därför skalas om till VIKT så att viktsumman stämmer med antalet reella observationer 2452.

```
160 PROC FREQ;
170   TABLES KV*ALD*UTB*POL*A/
180          LIST OUT=TABELL;
190   WEIGHT=VIKT;
```

Variabeln VIKT ger här frekvensen i den vägda syntetiska kontingenstabellen. LIST ger här listformat på kontingenstabellen, i annat fall skulle en mängd utskrifter erhållas.
I den skapade frekvenstabellen (som här heter TABELL) har alltid variabeln som anger frekvensen namnet COUNT vilket åberopas nedan på rad 210.

```
200 PROC FUNCAT DATA=TABELL;
210   WEIGHT=COUNT;
220   MODEL A=KV ALD UTB POL UTB*ALD
230     POL*KV POL*ALD UTB*POL /
240     FREQ ONEWAY CORRB P X;
250   RESPONSE 0 1;
```

Om rad 250 utesluts erhålles automatiskt en logitmodell. Här betyder rad 250 att den beroende variabeln utgöres av 0·(andelen med lägsta A-värdet)+ 1·(andelen med högsta A-värdet). Högsta A-värdet är 1 vilket anger politisk aktivitet alltså utgörs be-

roende variabeln av "andelen politiskt aktiva". (Deklarationen RESPONSE 1 0; hade givit "andelen ej politiskt aktiva" som beroende variabel.)

## 10.5 FUNCAT: Resultatutskrift

Identifiering. Vi skall studera de delar av utskriften som kan vara svåra att tolka. Ett stycke ned på utskriften återfinnes cellnumreringen (SAMPLE) vilken utgör den nyckel som behövs längre ned för att tyda designvariablerna. De två första raderna av beskrivningen av de 24 cellerna ser ut som följer:

| | | DESIGN | | | RESPONSE FREQUENCIES | | TOTAL |
|---|---|---|---|---|---|---|---|
| SAMPLE | KV | ALD | UTB | POL | 1 | 2 | |
| 1 | 1 | 0 | 1 | 1 | 63 | 11 | 73.4 |
| 2 | 1 | 0 | 1 | 5 | 192 | 26 | 217.7 |

Koderna 1 och 2 för RESPONSE FREQUENCIES anger klassnummer. Lägsta A-värdet som är 0 har klassnummer 1, det andra A-värdet som är 1 har klassnummer 2. Frekvenserna i dessa klasser är av programmet avrundade till heltal varför summan ej stämmer exakt med TOTAL vilken ges med 1 decimal. Direkt efter cellbeskrivningen följer designmatrisen. De två första raderna för cellerna 1 och 2 ges nedan:

| | | DESIGN MATRIX | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| SAMPLE | RESPONSE FUNCTION | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| 1 | 0.147246 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 1 |
| 2 | 0.118023 | 1 | 1 | 1 | 0 | 1 | -1 | 1 | 0 | -1 | -1 | 0 | -1 |

Designmatrisen bestämmer modellens parametrisering med utgångspunkt från MODEL-satsen och variabelkodningen i datamängden GRUND. Principerna beskrevs i kap. 10.2 och 10.3. Designvariablerna refererar till följande modelltermer:

| Designvariabel nr | Egenskap |
|---|---|
| 1 | Konstant term |
| 2 | KV |
| 3, 4 | ALD |
| 5 | UTB |
| 6 | POL |
| 7, 8 | UTB*ALD |

| Designvariabel nr | Egenskap |
|---|---|
| 9 | POL*KV |
| 10, 11 | POL*ALD |
| 12 | UTB*POL |

Modelltest. Testprocedurerna vid modellbyggnaden liknar i stora drag regressionsanalysens. Testen genomförs emellertid med en testvariabel som är asymptotiskt $\chi^2$-fördelad. Testmetodiken är följaktligen applicerbar för stora urval. Först testas om modellen på ett adekvat sätt beskriver observerade andelen politiskt aktiva inom cellerna. Testvariabeln utgörs av en vägd summa av kvadrerade avvikelser mellan observerad och predikterad proportion politiskt aktiva. Nollhypotesen kan uttryckas "Modellen passar". Tabell 10.5 visar att observerade proportioner inte signifikant avviker från predikterade förutsatt att testnivån är 0.20 eller lägre. Sannolikheten för ett $\chi^2$-värde större än eller lika med det observerade för RESIDUAL är nämligen nominellt 0.23. Här gäller liksom tidigare i rapporten att testproceduren ej tar hänsyn till att det är fråga om ett komplext urvalsförfarande.
Raderna i tabell 10.5 ger olika marginella $\chi^2$-test. För varje term testas nollhypotesen att denna terms parameter (de två parametrarna om ALD ingår i termen) är 0 under förutsättning att övriga termer ingår i modellen. Huvudeffekten för en faktor är inte meningsfullt att testa om denna faktor har samspelstermer som ingår i modellen. Om testen genomföres på den nominella nivån 0.05 är det endast samspelstermen UTB*POL som är insignifikant. Precis samma resultat erhölls för regressionsmodellen 5 som hade samma väntevärdesfunktion som modell 11. I det fallet behölls samspelstermen med motiveringen att det rimligen bör finnas ett samspel mellan UTB och POL, ett argument som självfallet är lika aktuellt i detta fall.

Tabell 10.5. $\chi^2$-test av MODELL 11 enligt FUNCAT.

| SOURCE | DF | CHI-SQUARE | PROB |
|---|---|---|---|
| Intercept | 1 | 361.30 | 0.0001 |
| KV | 1 | 26.01 | 0.0001 |
| ALD | 2 | 41.73 | 0.0001 |
| UTB | 1 | 14.35 | 0.0002 |
| POL | 1 | 34.25 | 0.0001 |
| ALD*UTB | 2 | 8.07 | 0.0177 |
| KV*POL | 1 | 9.10 | 0.0026 |
| ALD*POL | 2 | 9.05 | 0.0108 |
| UTB*POL | 1 | 3.24 | 0.0718 |
| RESIDUAL | 12 | 15.15 | 0.2334 |

Modell 11 har jämförts med större modeller som skiljer sig från modell 11 på följande sätt:

MODELL 12: Följande 2 termer tillkommer

| | CHI-SQUARE | PROB |
|---|---|---|
| KV*ALD | 3.55 | 0.1692 |
| KV*UTB | 0.04 | 0.8397 |

(KV*UTB stryks ur modellen. Eftersom $\chi^2$-testet här är marginellt måste sedan KV*UTB utvärderas genom att ställa MODELL 13 mot MODELL 11.)

MODELL 13: KV*ALD tillkommer

| | CHI-SQUARE | PROB |
|---|---|---|
| KV*ALD | 3.72 | 0.1559 |

(KV*ALD stryks vilket ger modell 11.)

För säkerhets skull undersöks om KV*UTB blir signifikant om KV*ALD saknas i modellen:

MODELL 14: KV*UTB tillkommer

| | CHI-SQUARE | PROB |
|---|---|---|
| KV*UTB | 0.21 | 0.6475 |

(KV*UTB strykes vilket ger modell 11.)

Således väljs modell 11 som slutmodell. Parameterskattningarna ges i tabell 10.6. Parametrarna är givna som avvikelser från medelvärdet och ej jämförbara med parametrarna i modell 5

förrän efter reparametrisering.

Reparametrisering. Det erbjuder ingen principiell svårighet att räkna fram värden för de parametrar som definierades i kapitel 8. Det är däremot mycket lätt att göra slarvfel eftersom varje sådan parameter erhålles som en summa av ett flertal avvikelseparametrar multiplicerade med 1 eller-1. Den önskade parametriseringen kan erhållas direkt med FUNCAT om man själv definierar sina designvariabler. Man använder då de definitioner av förklaringsvariabler (0-1 variabler) som utnyttjades i kapitel 8. Förändringarna i FUNCAT-programmet består i att dessa variabler måste bildas i datamängden GRUND och sedan deklareras i en DIRECT-sats, (vilken säger att indikatorvariablernas verkliga värden skall användas) samt att den kontingenstabell som skapas får flera dimensioner, en för varje indikatorvariabel. MODEL-satsen ändras i konsekvens härmed.

Rad 170 blir då

```
170  TABLES KV*ALDM*ALDH*UTB*POL*A/
```

Före MODEL-satsen inskjuts en rad

```
215  DIRECT KV ALDM ALDH UTB POL;
```

Resultat. Parameterskattningar ges i tabell 10.6 i avvikelseform och i tabell 10.7 i samma form som i kapitel 8 med yngre outbildade män utan politisk engagemang i familjen som referensgrupp. Nivån i denna grupp representeras av interceptet.

MODELL 5 och MODELL 11 ger ganska likartade parameterskattningar. GLS-skattningen av MODELL 11 leder dock till genomsnittligt lägre medelfel. (Tabell 10.7)
De båda modellernas prediktioner för olika celler är också snarlika. Genomsnittlig absolut prediktionsfel eller på engelska "Mean Absolute Deviation" (MAD) är praktiskt taget lika stora för MODELL 5 och MODELL 11. Inga större skillnader kan heller iakttagas mellan skattningen av variablernas effekter för olika delgrupper (celler) vilka redovisas i texten efter tabell 10.8.
Sammanfattningsvis kan konstateras att GLS-skattning av en linjär modell för kategoriska variabler vanligen är att föredra framför OLS-skattning på grund av att medelfelen för parameterskattningarna brukar bli mindre.

Tabell 10.6. Parameterskattningar i avvikelseform för MODELL 11 enligt FUNCAT. (Resultatens osäkerhet underskattas p.g.a. urvalsform och bortfall.)

| Effekt | Parameter | Df | Param.-est. | $\chi^2$ | Sannol. | Medelfel |
|---|---|---|---|---|---|---|
| INTERCEPT | 1 | 1 | 0.198 | 361.30 | 0.0001 | 0.010 |
| KV | 2 | 1 | -0.046 | 26.01 | 0.0001 | 0.009 |
| ALD | 3 | 1 | -0.084 | 40.69 | 0.0001 | 0.013 |
| | 4 | 1 | 0.032 | 5.28 | 0.0216 | 0.014 |
| UTB | 5 | 1 | 0.042 | 14.35 | 0.0002 | 0.011 |
| POL | 6 | 1 | 0.058 | 34.25 | 0.0001 | 0.010 |
| ALD*UTB | 7 | 1 | -0.013 | 1.96 | 0.1620 | 0.009 |
| | 8 | 1 | -0.022 | 4.59 | 0.0322 | 0.010 |
| KV*POL | 9 | 1 | -0.027 | 9.10 | 0.0026 | 0.009 |
| ALD*POL | 10 | 1 | -0.027 | 4.10 | 0.0429 | 0.013 |
| | 11 | 1 | 0.037 | 7.25 | 0.0071 | 0.014 |
| UTB*POL | 12 | 1 | -0.019 | 3.24 | 0.0718 | 0.011 |

Tabell 10.7. Parameterskattningar med 0-1 definierade förklaringsvariabler. MODELL 5 och MODELL 11. (Resultatens osäkerhet underskattas p.g.a. urvalsform och bortfall.)

| Variabel | MODELL 5 | | MODELL 11 | |
|---|---|---|---|---|
| | Param. est. | Medelfel | Param. est. | Medelfel |
| Intercept | 0.061 | 0.025 | 0.054 | 0.014 |
| KV | -0.042 | 0.017 | -0.038 | 0.014 |
| ÅLDM | 0.056 | 0.030 | 0.060 | 0.020 |
| ÅLDH | 0.068 | 0.030 | 0.071 | 0.020 |
| UTB | 0.089 | 0.028 | 0.096 | 0.019 |
| UTB*ÅLDM | -0.011 | 0.037 | -0.018 | 0.031 |
| UTB*ÅLDH | 0.101 | 0.041 | 0.098 | 0.039 |
| POL | 0.151 | 0.045 | 0.154 | 0.048 |
| POL*ÅLDM | 0.142 | 0.041 | 0.129 | 0.043 |
| POL*ÅLDH | 0.047 | 0.049 | 0.033 | 0.053 |
| POL*KV | -0.116 | 0.035 | -0.110 | 0.036 |
| POL*UTB | -0.066 | 0.039 | -0.077 | 0.043 |

Tabell 10.8. Prediktion av politisk aktivitet genom partier med MODELL 5 och MODELL 11. Grupperna kodas i enlighet med tabell 8.9.

| Grupp KÅUP | Obs frekv. (vägd) | MODELL 5 Pred. | MODELL 5 Resid. | MODELL 11 Pred. | MODELL 11 Resid. |
|---|---|---|---|---|---|
| 1000 | 0.022 | 0.019 | 0.003 | 0.016 | 0.006 |
| 0000 | 0.042 | 0.061 | -0.019 | 0.054 | -0.012 |
| 1001 | 0.060 | 0.054 | 0.006 | 0.061 | -0.001 |
| 1100 | 0.071 | 0.075 | -0.004 | 0.077 | -0.006 |
| 1200 | 0.078 | 0.087 | -0.009 | 0.087 | -0.009 |
| 1010 | 0.117 | 0.108 | 0.009 | 0.112 | 0.006 |
| 0100 | 0.121 | 0.117 | 0.004 | 0.114 | 0.007 |
| 1201 | 0.143 | 0.169 | -0.026 | 0.165 | -0.041 |
| 0200 | 0.142 | 0.129 | 0.013 | 0.125 | 0.017 |
| 1011 | 0.147 | 0.077 | 0.070 | 0.080 | 0.067 |
| 0010 | 0.148 | 0.150 | -0.002 | 0.150 | -0.002 |
| 0011 | 0.158 | 0.235 | -0.077 | 0.227 | -0.069 |
| 1110 | 0.160 | 0.153 | 0.007 | 0.154 | 0.005 |
| 1111 | 0.167 | 0.264 | -0.103 | 0.250 | -0.083 |
| 0110 | 0.189 | 0.195 | -0.006 | 0.192 | -0.003 |
| 1210 | 0.247 | 0.277 | -0.030 | 0.281 | -0.033 |
| 1101 | 0.270 | 0.255 | 0.018 | 0.250 | 0.020 |
| 0001 | 0.281 | 0.212 | 0.069 | 0.209 | 0.072 |
| 0210 | 0.321 | 0.319 | 0.002 | 0.318 | 0.003 |
| 0201 | 0.325 | 0.327 | -0.002 | 0.313 | 0.012 |
| 1211 | 0.351 | 0.293 | 0.058 | 0.281 | 0.070 |
| 0101 | 0.390 | 0.410 | -0.020 | 0.398 | -0.008 |
| 0211 | 0.458 | 0.451 | 0.007 | 0.429 | 0.029 |
| 0111 | 0.473 | 0.422 | 0.051 | 0.398 | 0.075 |
| MAD* | | | 0.026 | | 0.027 |

*MAD=Mean absolute deviation=$\frac{1}{n}\Sigma|\text{Obs-Pred}|$

MODELL 11:s "effekter" presenteras nedan vid sidan av "effekterna" för MODELL 5 hämtade från kapitel 8.12.

1. Predikterade sannolikheter för personer med enbart folkskola/grundskola och utan politiskt engagemang i familjen under uppväxttiden.

| | MODELL 5 | | | MODELL 11 | | |
|---|---|---|---|---|---|---|
| | Yngre | Medelålders | Äldre | Yngre | Medelålders | Äldre |
| Män | 0.061 | 0.117 | 0.129 | 0.054 | 0.114 | 0.125 |
| Kv. | 0.019 | 0.075 | 0.087 | 0.016 | 0.077 | 0.087 |

2. Utbildningseffekter då politiskt engagemang i familjen saknas.

| MODELL 5 | | | MODELL 11 | | |
|---|---|---|---|---|---|
| Yngre | Medelålders | Äldre | Yngre | Medelålders | Äldre |
| 0.089 | 0.078 | 0.190 | 0.096 | 0.077 | 0.194 |

3. Effekten av politiskt engagemang i familjen för personer med enbart folkskola/grundskola.

| | MODELL 5 | | | MODELL 11 | | |
|---|---|---|---|---|---|---|
| | Yngre | Medelålders | Äldre | Yngre | Medelålders | Äldre |
| Män | 0.151 | 0.293 | 0.198 | 0.155 | 0.284 | 0.188 |
| Kv. | 0.035 | 0.177 | 0.082 | 0.045 | 0.173 | 0.078 |

4. <u>Kombinerad</u> effekt av utbildning över folkskola/grundskola och politiskt engagemang i familjen.

| | MODELL 5 | | | MODELL 11 | | |
|---|---|---|---|---|---|---|
| | Yngre | Medelålders | Äldre | Yngre | Medelålders | Äldre |
| Män | 0.174 | 0.305 | 0.322 | 0.173 | 0.284 | 0.304 |
| Kv. | 0.058 | 0.189 | 0.206 | 0.062 | 0.173 | 0.194 |

Som synes är skillnaderna små mellan de båda modellernas skattningar av "effekter". Största differensen uppgår till 0.021 för medelålders män under punkt 4.

## 10.6 Jämförelse av vägda och ovägda skattningar

Kvoten mellan lägsta inklusionssannolikhet för personer (Uppsala) och högsta inklusionssannolikhet (Grästorp) är cirka 1:200. Vid de hittills genomförda modellskattningarna i denna rapport har det vägts med hänsyn till dessa sannolikheter.
Skattningarna har antingen baserats på enskilda individobservationer vilka vägts med inverterade inklusionssannolikheter (korrigerade för svarsbortfall) eller på en kontingenstabell som erhållits från rådata medelst en motsvarande vägningsprocedur.
En sådan vägningsprocedur eliminerar systematiska fel i skattningen av modellparametrar men de använda datorprogrammen ger ej en korrekt skattning av varianser varför testprocedurerna ej blir adekvata.

För att illustrera vägningens effekter har MODELL 5 och MODELL 11 som har samma väntevärdesfunktion med olika variansstruktur skattats även utan vägning med inverterade inklusionssannolikheter.
Tabell 10.9 visar att de båda ovägda skattningarna av modellerna ger snarlika resultat men att dessa skiljer sig mycket starkt från de båda vägda skattningarna vilka som tidigare visats sinsemellan också är snarlika. Den allra största skillnaden uppträder för termen POL*ALDM med skattningen 0.002 för båda ovägda skattningarna mot 0.142 (MODELL 5) resp. 0.139 (MODELL 11) för de vägda skattningarna. Det är uppenbart att för ett urval med så varierande inklusionssannolikheter som det här är fråga om är ovägda skattningar helt otillfredsställande.

Tabell 10.9. Skattningar med resp. utan vägning för skilda inklusionssannolikheter. MODELL 5 och MODELL 11 med 0-1 definierade förklaringsvariabler.

| Effekt | Vägt | | Ovägt | |
|---|---|---|---|---|
| | MODELL 5 | MODELL 11 | MODELL 5 | MODELL 11 |
| INTERCEPT | 0.061 | 0.054 | 0.059 | 0.054 |
| KV | -0.042 | -0.038 | -0.044 | -0.039 |
| ALDM | 0.056 | 0.060 | 0.116 | 0.117 |
| ALDH | 0.068 | 0.071 | 0.098 | 0.100 |
| UTB | 0.089 | 0.096 | 0.066 | 0.070 |
| UTB*ALDM | -0.011 | -0.018 | -0.057 | -0.061 |
| UTB*ALDH | 0.101 | 0.098 | 0.127 | 0.123 |
| POL | 0.151 | 0.154 | 0.209 | 0.208 |
| POL*ALDM | 0.142 | 0.129 | 0.002 | 0.002 |
| POL*ALDH | 0.047 | 0.033 | 0.028 | 0.018 |
| POL*KV | -0.116 | -0.110 | -0.109 | -0.104 |
| POL*UTB | -0.066 | -0.077 | -0.068 | -0.068 |

# 11. MULTIPLIKATIVA MODELLER

## 11.1 Symmetrisk modell

Låt säga att man med hjälp av en survey vill studera hur inställningen till löntagarfonder varierar med kön, ålder, partitillhörighet och fackförbundstillhörighet. Data från en sådan studie sammanställs i en kontingenstabell. Förhållandet kan beskrivas med hjälp av en asymmetrisk modell enligt vilken något lämpligt mått på inställningen till fonder (såsom andelen positiva eller logit för andelen positiva) beskrivs med en linjär modell i vilken kön, ålder, partitillhörighet och fackförbundstillhörighet utgör förklaringsvariabler. Sådana modeller har behandlats i de föregående kapitlen.

Det är också möjligt att använda en symmetrisk modell i vilken alla variabler behandlas symmetriskt som faktorer. Modellen förklarar då hur antalet observationer i cellerna i en 5-dimensionell kontingenstabell varierar med de 5 faktorerna. Fördelen med denna modellansats ligger i att man lättare kan studera högre grader av samspel samt att man kan belysa alla ömsesidiga kopplingar mellan faktorer, inte bara kopplingarna mellan fondinställning och övriga variabler som kan studeras med en asymmetrisk modell. Den senare fördelen är mera uppenbar i situationer i vilka man samtidigt studerar flera attitydvariabler, inte bara en som i exemplet ovan.

## 11.2 Produktform och logaritmerad form

Modellen för det symmetriskt formulerade problemet är i sin grundform en produktmodell. För att göra sådana mera lätthanterliga vid estimation brukar de logaritmeras. Detta ger en linjär modell i vilken termerna består av logaritmer. Tyvärr uppfattas denna modellform som mera svårförståelig än den välkända variansanalysmodellen med vilken den i vissa avseenden företer slående likheter (i logaritmerad form) medan den i andra skiljer sig markant från denna. Framställningar av symmetriska modeller brukar i allmänhet avse den loglinjära formen. Detta gäller exempelvis samtliga referenser i kapitel 9.2 utom Knoke and Burke [19].

Den fortsatta framställningen i föreliggande rapport kommer att avse produktformen av den symmetriska modellen. Parametrarna i produktformen kan tolkas i termer av geometriska medelvärden och odds.

## 11.3 Geometriskt medelvärde

Vi är vana vid att så snart det är fråga om medelvärden använda aritmetiska medelvärden. Det aritmetiska medelvärdet av n observationer uppfattas som ett enkelt och lättolkat begrepp.
Det geometriska medelvärdet definierat som n:te roten ur produkten av samtliga observationer

$$G = \sqrt[n]{X_1 X_2 \ldots X_n} = (X_1 X_2 \ldots X_n)^{\frac{1}{n}}$$

verkar däremot främmande och är ofta svårare att ge någon reell och lättförståelig innebörd.
Det aritmetiska och det geometriska medelvärdet skiljer sig för sneda fördelningar kraftigt från varandra.
I tabell 11.1 illustreras att aritmetiskt och geometriskt medelvärde i vissa situationer till och med kan vara av helt olika storleksordning.

Tabell 11.1. Geometriskt och aritmetiskt medelvärde.

| Observationer | | | | Geometriskt medelvärde | Aritmetiskt medelvärde |
|---|---|---|---|---|---|
| 4 | 4 | 4 | 4 | 4 | 4 |
| 2 | 4 | 4 | 8 | 4 | 5 |
| 2 | 2 | 4 | 16 | 4 | 6 |
| 1 | 2 | 8 | 16 | 4 | 6.75 |
| 1 | 1 | 16 | 16 | 4 | 8.5 |
| 2 | 2 | 2 | 32 | 4 | 9.5 |
| 1 | 2 | 2 | 64 | 4 | 17.25 |
| 1 | 1 | 2 | 128 | 4 | 33 |
| 1 | 1 | 1 | 256 | 4 | 64.75 |

## 11.4 Mättad modell för en 2x2-tabell

Vi startar med en modell för en 2x2-tabell med två kategoriska variabler på vardera två nivåer, säg kön (K) och partiblock (B).
Tabellen anger förväntat antal observationer $F_{ij}$ för cell (i,j) vid obundet slumpmässigt urval av n individer från populationen. (Tabell 11.2.)

$$F_{ij} = \mu\theta_i^K\theta_j^B\theta_{ij}^{KB} \qquad (11.1)$$

Tabell 11.2. Förväntade cellfrekvenser för en mättad modell.

| | | Partiblock: Socialistiskt (1) | Partiblock: Borgerligt (2) |
|---|---|---|---|
| Kön | Kvinnor (1) | $F_{11}=\mu\theta_1^K\theta_1^B\theta_{11}^{KB}$ | $F_{12}=\mu\theta_1^K\theta_2^B\theta_{12}^{KB}$ |
| | Män (2) | $F_{21}=\mu\theta_2^K\theta_1^B\theta_{21}^{KB}$ | $F_{22}=\mu\theta_2^K\theta_2^B\theta_{22}^{KB}$ |

Symboliskt kan modellen skrivas:

MODELL K B KB

I modell (11.1) är $\mu$ en allmän nivåparameter definierad som geometriska medelvärdet av samtliga cellers väntevärden.

$$\mu = \sqrt[4]{F_{11}F_{12}F_{21}F_{22}} = (F_{11}F_{12}F_{21}F_{22})^{1/4} \qquad (11.2)$$

Den fingerade väntevärdesfördelningen i tabell 11.3 ger

$$\mu = \sqrt[4]{900\cdot 400\cdot 100\cdot 400} = 346{,}41$$

Tabell 11.3. Fingerad väntevärdesfördelning med avseende på kön och partiblock.

| | | Block: Soc. | Block: Borg. | |
|---|---|---|---|---|
| Kön | Kv. | 900 | 400 | 1300 |
| | Män | 100 | 400 | 500 |
| | | 1000 | 800 | 1800 |

Huvudeffekten $\theta_i^K$, där K står för kön med i=1 för kvinnor

och i=2 för män, jämför geometriska medelfrekvensen för cellerna inom rad i med geometriska medelfrekvensen $\mu$ för hela tabellen. Ett värde på $\theta_1^K$ som är mindre än 1, lika med 1 respektive större än 1 skulle betyda att geometriska medelfrekvensen för kvinnor är lägre än, lika med respektive högre än referensnivån $\mu$.

Alltså

$$\theta_1^K = \frac{\sqrt{F_{11}F_{12}}}{4\sqrt{F_{11}F_{12}F_{21}F_{22}}} \qquad (11.3a)$$

$$\theta_2^K = \frac{\sqrt{F_{21}F_{22}}}{4\sqrt{F_{11}F_{12}F_{21}F_{22}}} \qquad (11.3b)$$

Som synes föreligger alltid restriktionen $\theta_1^K\theta_2^K = 1$ varför den ena parametern är överflödig. Det är tillräckligt att ange $\theta_1^K$ eftersom $\theta_2^K = 1/\theta_1^K$ .

Från rad 1 erhålles geometiska medelfrekvensen för kvinnor $\sqrt{F_{11}F_{12}}=\sqrt{900\cdot400}=600$ och således $\theta_1^K=600/346,41=1,732$ .
För män blir $\theta_2^K=1/1,732=0,577$ .

På motsvarande sätt definieras huvudeffekterna $\theta_1^B$ och $\theta_2^B$ som genomsnittliga kolumneffekter för socialistiska blocket respektive borgerliga blocket. Man erhåller geometriska medelvärdet $\sqrt{900\cdot100}=300$ för kolumn 1 och $\sqrt{400\cdot400}=400$ för kolumn 2.

Således $\theta_1^B = 300/346,41 = 0,866$ och $\theta_2^B = 400/346,41 = 1,155$ .
Geometriska medelfrekvensen för socialistiska blocket ligger således 13,4 % under referensnivån 346,41 trots att det finns betydligt flera socialistiska sympatisörer (1000) än borgerliga sympatisörer (800). Geometriska medelfrekvensen för borgerliga sympatisörer ligger 15,5 % över referensnivån. På grund av att vi är så vana vid att tänka i termer av aritmetiska medelvärden och totalantal blir således huvudeffekten för politiskt block direkt vilseledande i detta fall.

Samspelseffekten $\theta_{ij}^{KB}$ definieras som kvoten mellan förväntat värde $F_{ij}$ och förväntad nivå med hänsyn till referensnivå och huvudeffekter. Man kan visa att

$$\theta_{11}^{KB} = \theta_{22}^{KB} = \frac{1}{\theta_{12}^{KB}} = \frac{1}{\theta_{21}^{KB}} = \sqrt[4]{\frac{F_{11}F_{22}}{F_{12}F_{21}}} \qquad (11.4)$$

Korsproduktkvoten kan skrivas som kvoten mellan radoddsen $F_{11}/F_{12}$ och $F_{21}/F_{22}$ och även som kvoten mellan kolumnoddsen $F_{11}/F_{21}$ och $F_{12}/F_{22}$ d.v.s.

$$\frac{F_{11}F_{22}}{F_{12}F_{21}} = \frac{F_{11}}{F_{12}} \Big/ \frac{F_{21}}{F_{22}} = \frac{F_{11}}{F_{21}} \Big/ \frac{F_{12}}{F_{22}} \qquad (11.5)$$

Vid oberoende mellan faktorerna K och B är de båda radoddsen lika (även kolumnoddsen är lika). Således blir oddskvoterna 1 och därmed korsproduktkvoten 1.
Eftersom det räcker att ange en samspelsterm brukar denna betecknas $\theta^{AB}$ och avse cell (1,1). I det aktuella exemplet blir

$$\theta^{KB} = \sqrt[4]{\frac{900 \cdot 400}{400 \cdot 100}} = 1{,}732$$

och $1/\theta^{KB} = 1/1{,}732 = 0{,}577$

Samspelsmönstret är således följande:

| | | Block 1 | Block 2 |
|---|---|---|---|
| Kön | 1 | 1,732 | 0,577 |
| | 2 | 0,577 | 1,732 |

Observera att inom varje rad och varje kolumn blir alltid produkten av samspelstermerna 1.
Exemplet har illustrerat att parametervärdena som sådana är föga informativa och ibland direkt vilseledande. Däremot är denna typ av modeller väl lämpade för att man skall kunna testa hypoteser huruvida vissa typer av samspel förekommer.

Parametrarna kan också tolkas i termer av odds och oddskvoter (se referenserna i kap. 9.4).

## 11.5 Enkla modeller för en 2x2-tabell

Tills vidare förutsätts enkelt slumpmässigt urval av individer från en population.

1. Den mest regelbundna modellen för två dikotoma variabler har samma förväntade frekvenser i alla 4 cellerna.

Tabell 11.4. Modell utan huvudeffekter. Förväntade frekvenser vid OSU av 200 individer.

| | $B_1$ | $B_2$ | |
|---|---|---|---|
| $A_1$ | 50 | 50 | 100 |
| $A_2$ | 50 | 50 | 100 |
| | 100 | 100 | 200 |

Modellen skrivs

$$F_{ij} = \mu \ .$$

Modellen tjänar ofta som referensbas vid modellbyggnad men saknar i övrigt intresse.

2. Nästa modell har olika stora förväntade marginalfrekvenser för en faktor (säg A) men lika stora för den andra (B). (Tabell 11.5.)

Tabell 11.5. Modell med huvudeffekt endast för A. Förväntade frekvenser vid OSU av 200 individer.

| | $B_1$ | $B_2$ | |
|---|---|---|---|
| $A_1$ | 40 | 40 | 80 |
| $A_2$ | 60 | 60 | 120 |
| | 100 | 100 | 200 |

Faktorerna A och B är oberoende.
Modellen skrivs matematiskt

$$F_{ij} = \mu \cdot \theta_i^A$$

och symboliskt

MODELL: A

Referensnivån blir enligt (11.2) $\mu$=48,99 och huvudeffekterna för A blir enligt (11.3) $\theta_1^A$=0,816 och $\theta_2^A$=1,225 .

3. Härnäst följer den vanliga modellen för oberoende mellan A och B. (Se tabell 11.6.) De två föregående modellerna är specialfall med oberoende faktorer.

Tabell 11.6. Modell med huvudeffekter för A och B. (Oberoendemodell) Förväntade frekvenser vid OSU av 200 individer.

| | $B_1$ | $B_2$ | |
|---|---|---|---|
| $A_1$ | 44 | 36 | 80 |
| $A_2$ | 66 | 54 | 120 |
| | 110 | 90 | 200 |

De betingade sannolikheterna (eller ekvivalent oddsen) är lika radvis (odds 11/9) och kolumnvis (odds 2/3). Korsprodukten $F_{11}F_{22}/F_{12}F_{21}$ är lika med 1.

Modellen skrivs

$$F_{ij} = \mu \cdot \theta_i^A \theta_j^B \qquad (11.6)$$

och symboliskt

MODELL: A B

Referensnivån blir nu $\mu$ = 48,74 medan $\theta_1^A$=0,816, $\theta_2^A$=1,225 och $\theta_1^B$=1,106, $\theta_2^B$=0,904 .

## 11.6 Hierarkiska modeller

Normalt brukar man endast arbeta med hierarkiska modeller. Hierarkisk modell betyder i detta sammanhang att om en viss samspelseffekt, säg ABC, ingår i en modell så ingår även alla huvudeffekter och samspelseffekter av lägre ordning som kan bildas genom att man väljer ut en delmängd av faktorerna i den högre samspelseffekten.
ABC implicerar således att även A,B,C, AB, AC och BC ingår i den hierarkiska modellen. Modellen A B C AB AC BC ABC kan därför kortare skrivas [ABC] medan den ovan behandlade modellen för 2x2-tabellen skrivs [KB].
I exempelvis en tredimensionell tabell kan de observerade frekvenserna exakt återges med hjälp av skattningar av parametrarna i modellen [ABC]. Modellen kallas därför mättad.
Målet vid konstruktion av modeller för kontingenstabeller är att på ett adekvat sätt representera datastrukturerna med hjälp av en icke mättad modell som är så enkel som möjligt och som har få parametrar.

## 11.7 Tredimensionella kontingenstabeller

Tre kategoriska variabler A, B och C observeras. Förväntade cellfrekvenser $F_{ijk}$ anges i en tredimensionell kontingenstabell med rader, kolumner och skikt.
Summation över skikt (k) för en given rad (i) och kolumn (j) ger förväntat antal händelser $A_iB_j$ .
Om summationen över skikt (k) utföres för alla kombinationer av i och j erhålles den 2-dimensionella marginalfördelningen för A och B.
Det finns ytterligare två 2-dimensionella marginalfördelningar, för AC och för BC, vilka erhålles på motsvarande sätt genom summation kolumnvis respektive radvis i den 3-dimensionella tabellen.
De 3 endimensionella marginalfördelningarna för A, B och C erhålles genom summation över ett index i en 2-dimensionell marginalfördelning.
För varje C-nivå förekommer en betingad 2-dimensionell fördelning för A och B. På motsvarnade sätt finns betingade fördelningar för A och C resp. B och C.
En rad intressanta frågeställningar kan resas i anslutning till dessa olika typer av fördelningar, bl.a.

a. Under vilka förutsättningar kan den 3-dimensionella fördelningen
   - (i) rekonstrueras ur de 3 endimensionella marginalfördelningarna ?
   - (ii) rekonstrueras ur en 2-dimensionell marginalfördelning (exempelvis AB) och en endimensionell marginalfördelning (C) ?

b. Kan det saknas association i en 2-dimensionell marginalfördelning (exempelvis A, B) om det föreligger beroende i betingade fördelningar för A och B på olika C-nivåer ?

c. Kan det finnas association i en 2-dimensionell marginalfördelning (säg A, B) om det föreligger oberoende i de betingade fördelningarna för A och B på varje C-nivå ?

## 11.8 Modeller för tredimensionella kontingenstabeller

1. Ömsesidigt oberoende mellan A, B och C.
   Varje cellsannolikhet kan anges som produkten av sannolikheterna i de endimensionella marginalfördelningarna

$$P_{ijk} = P_{i..}P_{.j.}P_{..k} \qquad (11.7)$$

Förväntad cellfrekvens kan då skrivas

$$F_{ijk} = \mu \cdot \theta_i^A \theta_j^B \theta_k^C \qquad (11.8)$$

Symboliskt anges

MODELL: A B C.

En modell med enbart huvudeffekter inplicerar således oberoende faktorer. Detta gäller även specialfallet då en eller flera huvudeffekter saknas d.v.s. motsvarande marginalfördelningar är likformiga.

2. Multipelt oberoende mellan C och den bivariata variabeln (A, B).
Den tredimensionella sannolikhetsfördelningen bestäms här av den tvådimensionella marginalfördelningen för (A, B) och den endimensionella marginalfördelningen för C enligt

$$p_{ijk} = p_{ij} \cdot p_{\cdot\cdot k} \qquad (11.9)$$

Förväntad cellfrekvens kan då skrivas

$$F_{ijk} = \mu \theta_i^A \theta_j^B \theta_k^C \theta_{ij}^{AB} \qquad (11.10)$$

Symboliskt anges

MODELL: A B C AB

eller kortare

MODELL: [AB C]

Om A anger partisympati, B åldersklass och C kön har man ett samband mellan parti och åldersklass men detta samband är det samma för män och kvinnor.
Eftersom C är oberoende av den bivariata variabeln (A, B) är också C oberoende av variablerna A och B var för sig. (Se tabell 11.7.)

Tabell 11.7. Multipelt oberoende mellan kön (C) och parti/åldersfördelning (AB).

a. Beroende i de betingade AB-fördelningarna för <u>given C-nivå</u>. (Oddskvot 1/2 för båda C-nivåerna.)

Kvinnor ($C_1$)

| | $B_1$ | $B_2$ | |
|---|---|---|---|
| $A_1$ | 40 | 80 | 120 |
| $A_2$ | 120 | 120 | 240 |
| | 160 | 200 | 360 |

Män ($C_2$)

| | $B_1$ | $B_2$ | |
|---|---|---|---|
| $A_1$ | 50 | 100 | 150 |
| $A_2$ | 150 | 150 | 300 |
| | 200 | 250 | 450 |

b. Oberoende i betingade AC-fördelningar för <u>given B-nivå</u>. (Oddskvot 1 för båda B-nivåerna.)

$B_1$

| | $C_1$ | $C_2$ | |
|---|---|---|---|
| $A_1$ | 40 | 50 | 90 |
| $A_2$ | 120 | 150 | 270 |
| | 160 | 200 | 360 |

$B_2$

| | $C_1$ | $C_2$ | |
|---|---|---|---|
| $A_1$ | 80 | 100 | 180 |
| $A_2$ | 120 | 150 | 270 |
| | 200 | 250 | 450 |

c. Oberoende i betingade BC-fördelningar för <u>given A-nivå</u>. (Oddskvot 1 för båda A-nivåerna.)

$A_1$

| | $C_1$ | $C_2$ | |
|---|---|---|---|
| $B_1$ | 40 | 50 | 90 |
| $B_2$ | 80 | 100 | 180 |
| | 120 | 150 | 270 |

$A_2$

| | $C_1$ | $C_2$ | |
|---|---|---|---|
| $B_1$ | 120 | 150 | 270 |
| $B_2$ | 120 | 150 | 270 |
| | 240 | 300 | 540 |

d. Bivariata marginalfördelningar:
Association AB (Oddskvot 1/2)
Ingen association AC (Oddskvot 1)
Ingen association BC (Oddskvot 1)

| | $B_1$ | $B_2$ | |
|---|---|---|---|
| $A_1$ | 90 | 180 | 270 |
| $A_2$ | 270 | 270 | 540 |
| | 360 | 450 | 810 |

| | $C_1$ | $C_2$ | |
|---|---|---|---|
| $A_1$ | 120 | 150 | 270 |
| $A_2$ | 240 | 300 | 540 |
| | 360 | 450 | 810 |

| | $C_1$ | $C_2$ | |
|---|---|---|---|
| $B_1$ | 160 | 200 | 360 |
| $B_2$ | 200 | 250 | 450 |
| | 360 | 450 | 810 |

3. Betingat oberoende mellan A och B.
Antag nu istället att bland kvinnor är ålder och partisympatier oberoende. Partifördelningen är med andra ord den samma för de båda kvinnliga åldersklasserna.
Bland män är partifördelningen en annan än bland kvinnor men även bland män är ålder och partisympatier oberoende. Eftersom åldersfördelningarna (B) bland män och kvinnor är olika erhålles en association mellan kön (C) och parti (A) liksom mellan ålder (B) och kön (C).

Förväntad cellfrekvens kan skrivas

$$F_{ijk} = \mu \cdot \theta^A_i \theta^B_j \theta^C_k \cdot \theta^{AC}_{ik} \theta^{BC}_{jk} \qquad (11.11)$$

Symboliskt anges

MODELL: A B C AC BC

eller kortare

MODELL: [AC BC]

(Se tabell 11.8)

Tabell 11.8. Betingat oberoende mellan partisympati (A) och åldersklass (B).

a. Oberoende i betingade AB-fördelningar för given C-nivå. (Oddset 1)

Kvinnor $C_1$

| | $B_1$ | $B_2$ | |
|---|---|---|---|
| $A_1$ | 60 | 60 | 120 |
| $A_2$ | 120 | 120 | 240 |
| | 180 | 180 | 360 |

Män $C_2$

| | $B_1$ | $B_2$ | |
|---|---|---|---|
| $A_1$ | 40 | 200 | 240 |
| $A_2$ | 60 | 300 | 360 |
| | 100 | 500 | 600 |

b. Beroende i betingade AC-fördelningar för given B-nivå. (Oddskvot 3/4)

| | $B_1$ | | | $B_2$ | | |
|---|---|---|---|---|---|---|
| | $C_1$ | $C_2$ | | $C_1$ | $C_2$ | |
| $A_1$ | 60 | 40 | 100 | 60 | 200 | 260 |
| $A_2$ | 120 | 60 | 180 | 120 | 300 | 420 |
| | 180 | 100 | 280 | 180 | 500 | 680 |

c. Beroende i betingade BC-fördelningar för given A-nivå. (Oddskvot 5)

| | $A_1$ | | | $A_2$ | | |
|---|---|---|---|---|---|---|
| | $C_1$ | $C_2$ | | $C_1$ | $C_2$ | |
| $B_1$ | 60 | 40 | 100 | 120 | 60 | 180 |
| $B_2$ | 60 | 200 | 260 | 120 | 300 | 420 |
| | 120 | 240 | 360 | 240 | 360 | 600 |

d. Bivariata marginalfördelningar:

Association AB (Oddskvot 35/39)
Association AC (Oddskvot 3/4 )
Association BC (Oddskvot 5 )

| | $B_1$ | $B_2$ | | $C_1$ | $C_2$ | |
|---|---|---|---|---|---|---|
| $A_1$ | 100 | 260 | 360 | 120 | 240 | 360 |
| $A_2$ | 180 | 420 | 600 | 240 | 360 | 600 |
| | 280 | 680 | 960 | 360 | 600 | 960 |

| | $C_1$ | $C_2$ | |
|---|---|---|---|
| $B_1$ | 180 | 100 | 280 |
| $B_2$ | 180 | 500 | 680 |
| | 360 | 600 | 960 |

4. Modellen med alla möjliga tvåfaktorsamspel mellan A, B och C men inget trefaktorsamspel har väntevärdet givet av

$$F_{ijk} = \mu\theta_i^A\theta_j^B\theta_k^C\theta_{ij}^{AB}\theta_{ik}^{AC}\theta_{jk}^{BC}$$

Symboliskt skrivs

MODELL: A B C AB AC BC

eller kortare

MODELL: [AB AC BC]

Nackdelen med denna modell är att den inte kan ges någon tolkning i konventionella termer i form av oberoende, betingat oberoende och lika sannolikheter.

5. I den mest komplicerade modellen av förväntad cellfrekvens ingår en trefaktorterm.

$$F_{ijk} = \mu\theta_i^A\theta_j^B\theta_k^C\theta_{ij}^{AB}\theta_{ik}^{AC}\theta_{jk}^{BC}\theta_{ijk}^{ABC}$$

symboliskt skrivs

MODELL: A B C AB AC BC ABC

eller kortare

MODELL: [ABC]

Detta är en mättad modell eftersom varje möjlig en-, två- och trefaktorterm ingår. Den mättade modellen ger fullständig anpassning till varje tänkbar observerad kontingenstabell.

Trefaktorsamspelstermen $\theta_{ijk}^{ABC}$ som är definierad för var och en av de 8 cellerna i en 2x2x2-tabell är liksom övriga $\theta$-termer föremål för restriktioner.
För varje C-nivå är produkten inom varje rad och varje kolumn lika med 1. Samma restriktion gäller även varje vertikal "pelare" av celler.
Alla 8 trefaktorsamspel kan uttryckas i en enda parameter $\theta^{ABC}$ som alltid avser cell (1,1,1). Strukturen framgår av följande figur.

| | $C_1$ | |
|---|---|---|
| | $B_1$ | $B_2$ |
| $A_1$ | $\theta^{ABC}$ | $1/\theta_{ABC}$ |
| $A_2$ | $1/\theta^{ABC}$ | $\theta^{ABC}$ |

| | $C_2$ | |
|---|---|---|
| | $B_1$ | $B_2$ |
| $A_1$ | $1/\theta^{ABC}$ | $\theta^{ABC}$ |
| $A_2$ | $\theta^{ABC}$ | $1/\theta^{ABC}$ |

Vid konstruktion av modeller för tredimensionella kontingenstabeller önskar man om möjligt beskriva fördelningen med någon av modellerna 1, 2 och 3 eftersom dessa som ovan visats medger en tolkning av sambanden i enkla och konventionella termer.

## 11.9 Felaktig analysmetod

Vid analys av multivariata samband studeras ofta en rad korstabeller för två variabler i taget. I många fall genomförs $\chi^2$-test av oberoende för dessa marginaltabeller. Förfaringssättet är helt felaktigt och leder ofta till falska slutsatser. Detta illustreras bl.a. av tabell 11.8 i vilken A och B är oberoende inom varje C-nivå. Trots detta föreligger association mellan A och B (oddskvot skild från 1) i marginalfördelningen AB. Det kan exempelvis också förekomma stark positivt beroende mellan A och B på båda C-nivåerna (oddskvoter större än 1) och ändå blir negativ association (oddskvot mindre än 1) i marginaltabellen AB. (För exempel se Upton [27] sid 43.)

Ett exempel: Serviceutnyttjande. I en av delstudierna i forskningsprogrammet för utvärdering av kommunreformen undersöker Olander och Widberg [23] hushållens utnyttjande av olika former av kommunal service. Författarna betraktar en bakgrundsfaktor i taget för att bestämma denna faktors effekt på serviceuttnyttjandet. På basis av denna analys rangordnas bakgrundsfaktorerna efter grad av påverkan på serviceutnyttjandet.

> Författarna kommenterar sitt förfaringssätt på följande sätt:
> "En svaghet är att tekniken inte ger direkta besked om i vilken utsträckning som faktorernas rangordning beror på deras egna förklaringsvärden och i vilken utsträckning som rangordningen beror på samvariation mellan vissa av faktorerna. Endast en multivariat analysteknik skulle kunna ge sådana direkta och kvantifierade besked. Problemet kan dock lösas genom att den använda enkla tekniken kompletteras med särskilda kommentarer kring samvariationen, baserade på logiska granskningar av bakgrundsmaterialet i dess helhet".

Författarna motsäger sig själva i citatet. Enda sättet att undvika hopblandning av olika variablers effekter är att analysera den samtidiga variationen i alla för sakproblemet relevanta variabler. En analys av ett antal två- eller tredimen-

sionella marginalfördelningar är otillräcklig vare sig bakgrundsvariablerna samspelar i sin inverkan på utnyttjandet av kommunal service eller ej. Själva existensen av samvariation mellan bakgrundsvariabler leder till hopblandning av effekter om analysen utgår från marginalfördelningar.

Ett exempel på hur man kan analysera Olander och Widbergs problem med produktmodeller ges längre fram i denna rapport.

## 11.10 Modeller för flerdimensionella tabeller

Hittills har bara 2x2 och 2x2x2-tabeller betraktats. Teorin kan emellertid lätt generaliseras till flera klasser per variabel och till tabeller med flera än 3 variabler. Generaliseringen av modellen är uppenbar och den allmänna tolkningen av modellerna är snarlik. Läsaren hänvisas till handbokslitteraturen. (Se referensen i kap. 9.4.)

## 11.11 Urvalsformer och experiment: Modellval

Teorin för produktmodellerna (de loglinjära modellerna) är utvecklad för tre olika samplingmodeller.

1. Poisson. Denna modelltyp hör huvudsakligen hemma i experimentella tillämpningar. Man observerar oberoende Poissonprocesser, över en på förhand bestämd observationsperiod. För varje cell betraktas frekvensen som en observation från en speciell sådan process. (Se Bishop et.al [1 ].)

2. Multinomial. Ett obundet slumpmässigt urval (OSU) av element väljs och klassificeras i en kontingenstabell.

3. Produktmultinomial. Ett stratifierat urval med obundet slumpmässigt urval inom strata väljs. För varje stratum klassificeras elementen i en kontingenstabell. I detta fall måste alltid stratifieringsvariabeln ingå i produktmodellen vare sig den är av direkt intresse eller ej.

För givna marginaler leder de tre modellerna till samma maximumlikelihoodestimatorer av förväntade cellfrekvenser.

## 11.12 Komplexa urval

Få surveys är genomförda med ett OSU av individer eller ett stratifierat urval med OSU inom strata. Frågan blir därför hur data erhållna enligt andra urvalsformer kan analyseras. Teoriutvecklingen härför är ännu ytterst begränsad.
I brist på teori som kan ge ledning verkar följande förfaringssätt vara rimliga nödfallsåtgärder förutsatt att resultaten tolkas försiktigt.

a. Självvägande urval enligt olika designs.
Dessa bör hjälpligt kunna analyseras enligt multinomialmodellen. Om endast ett fåtal förstastegsenheter utvalts bör man överväga att analysera enligt produktmultinomialmodellen, d.v.s. införa förstastegsenhet som en faktor i modellen.

b. Icke självvägande urval (bortsett från stratifierat urval med OSU inom strata).

(i) Fåtal förstastegsenheter i urvalet.
Inom varje förstastegsenhet beräknas en viktad kontingenstabell med inverterade inklusionssannolikheter som vikter. Vikterna standardiseras så att de vägda frekvenserna summeras till den verkliga urvalsstorleken. (Om urvalet inom en förstastegsenhet är självvägande bortfaller viktningen.)
Produktmultinomialmodellen används, d.v.s. förstastegenhetstillhörighet införes som en faktor i modellen.

(ii) Ett flertal förstastegsenheter i urvalet.
En viktad frekvenstabell beräknas för hela urvalet. Multinomialmodellen används.

# 12. MODELLBYGGNAD - POLITISK AKTIVITET

## 12.1 Mål och arbetssätt

Vid modellkonstruktion söker man sig successivt fram till en adekvat modell. Det är inte ovanligt att man under arbetets gång studerar ett tiotal modeller och ibland betydligt flera. De stegvisa sökprocedurerna är i princip de samma som vid regressionsanalys. Man startar antingen med en enkel modell som byggs ut succesivt eller med den mättade modellen som succesivt reduceras.

Det förekommer ett flertal datorprogram som utnyttjar olika beräkningsalgoritmer varför skattningarna troligen ej blir identiska med två olika program.

Analysen förutsätter som tidigare nämnts en hierarkisk modell.

Målet är att finna en enkel modell med få komponenter som på ett adekvat sätt kan förklara datastrukturen. Det finns ofta två eller flera ungefär lika enkla modeller som ger en god beskrivning.

Man vill helst undvika modeller med samspel mellan fler än 3 variabler. Skälet är att man eljest dels får många termer i modellen (ett samspel av högre ordning drar med sig en hel rad termer eftersom modellen skall vara hierarkisk), dels får svårt att tolka modellen och dels kan få ett lågt frihetsgradtal.

Stora tabeller med många celler innehåller vanligen många tomma celler vilket resulterar i att modellen inte kan skattas på ett godtagbart sätt. Det är därför angeläget att endast ha ett fåtal klasser för varje variabel. Om klassindelningen blir meningsfull är det en fördel med bara 2 eller 3 klasser. Man konstruerar ogärna modeller för flera än 4 à 5 variabler.

## 12.2 Testprocedurer

Låt $f_{ijk}$ ange observerat antal individer i cell (i,j,k) och $F_{ijk}$ ange förväntat antal individer i cellen under en viss modell. Modellens anpassning kan testas med Pearsons klassiska $\chi^2$-test eftersom testvariabeln

$$X^2 = \sum_{i,j,k} \frac{(f_{ijk}-F_{ijk})^2}{f_{ijk}} \qquad (12.1)$$

asymptotiskt är $\chi^2$-fördelad med frihetsgradtalet n-p där

n = antal celler och p = antal parametrar i modellen.
Om anpassningen är god d.v.s. skillnaden mellan observerade och förväntade frekvenser är liten blir det observerade $\chi^2$-värdet lågt. Som kriterium på acceptabel modell har man följande: Modellen godtas om sannolikheten att erhålla en större avvikelse mellan observerad och förväntad frekvens än den som urvalet uppvisar är större än ett förutbestämt tal $\alpha$ . Man godtar alltså modellen om testet inte ger signifikant utslag.
Ett annat test av samma hypotes ges av likelihood-ratiostorheten

$$G^2 = 2 \sum_{i,jik} f_{ijk} \ln(f_{ijk}/F_{ijk}) \qquad (12.2)$$

som också är asymptotiskt $\chi^2$-fördelad med frihetsgradtalet n-p .
Likelihood-ratio variabeln $G^2$ har en stor fördel framför det klassiska avvikelsemåttet $X^2$. $G^2$ är nämligen i motsats till $X^2$ additivt vid uppdelning av termer som tillhör två modeller $M_1$ och $M_2$, sådana att parametrarna i $M_1$ är en delmängd av parametrarna i $M_2$. Skillnaden mellan $G^2$-måttet för modellerna ger ett test av hypotesen att de extra parametrarna i modell $M_2$ är 0 (modellerna är angivna i linjär eller loglinjär form) givet parametrarna i $M_1$.
Denna differens är asymptotiskt $\chi^2$-fördelad med ett frihetsgradtal som är lika med differensen mellan frihetsgradtalen för $M_2$ och $M_1$. Om detta test ger signifikans, d.v.s. om sannolikheten att erhålla en differens lika stor som eller större än den observerade är mindre än ett givet $\alpha$ förkastas nollhypotesen att extraparametrarna i modell $M_2$ är 0. Således underkänns modell $M_1$. Modellen $M_2$ är acceptabel om anpassningstestet för $M_2$ inte ger signifikans.
Med hjälp av likelihood-ratio test för en följd av modeller och test av de parametrar som konstituerar skillnaden mellan olika par av modeller söker man sig fram till en eller ett fåtal modeller som står i överensstämmelse med de observerade cellfrekvenserna. Därpå återstår att med hjälp av ytterligare hjälpmedel välja den "bästa" modellen.

## 12.3 Determinationsmått

Inom multipel linjär regressionsanalys används multipla determinationskoefficienten (multipla korrelationskoefficienten i kvadrat: $R^2$) som mått på hur bra prediktion en viss modell ger. $R^2_{y \cdot x_1 x_2}$ anger hur stor den relativa minskningen av den oförklarade variationen i y är om man predikterar y-värden med hjälp av variablerna $x_1$ och $x_2$ jämfört med fallet då man inte använder någondera av $x_1$ och $x_2$.

På motsvarande sätt anger partiella korrelationskoefficienten i kvadrat $r^2_{yx_2 \cdot x_1}$ hur stor den relativa minskningen av den oförklarade variansen blir om y predikteras med hjälp av både $x_1$ och $x_2$ istället för att prediktera $y^2$ med hjälp av enbart $x_1$.

Goodman [ 8], [ 9] definierar liknande mått för analys av kvalitativa variabler. Man utgår från en lämpligt vald basmodell. I studier av det symmetriska fallet med enbart faktorer utgörs denna vanligen av modellen med enbart enfaktorstermer. Med 5 faktorer A,B,C,D och E blir modellen med oberoende mellan samtliga variabler d.v.s. [A B C D E] basmodell. Vid studier av en responsvariabel och dess beroende av ett antal faktorer utgör modellen [A,BCDE] som antar oberoende mellan responsvariabeln A och faktorerna B,C,D och E (vilka däremot inbördes förutsättes beroende) lämplig basmodell.

Som mått på oförklarad variation används antingen Pearson's $\chi^2$ definierat enligt (12.1) eller informationsmåttet $G^2$ ("maximumlikelihood $\chi^2$") definierat enligt (12.2).
Om basmodellens parametrar utgör en delmängd av den alternativa modellens parametrar difinieras den senares multipla determination som

$$R^2 = \frac{\chi^2_{bas} - \chi^2_{alt.}}{\chi^2_{bas}} \qquad (12.3)$$

Definitionen av $R^2$ för informationsmåttet $G^2$ är identisk, man byter således $\chi^2$ mot $G^2$ i (12.3).

Vid jämförelse av två modeller A1 och A2 som skiljer sig åt genom att A2 tillförts en parameter (säg $\theta_{12}$) som ej ingår i $A_1$ definieras den partiella determinationen som

$$r^2_{f\theta_{12}\cdot A_1} = \frac{\chi^2_{A1} - \chi^2_{A2}}{\chi^2_{A1}} \qquad (12.4)$$

Indexet f står här för observerad cellfrekvens i kontingenstabellen. Liksom i den multipla regressionsanalysen beror den partiella determination som en viss parameter ger av vilka andra parametrar som tidigare ingår i modellen. Det är därför fel att säga att en viss parameter "förklarar" x procent av variationen. Det korrekta uttrycket är: Parametern (säg $\theta_{12}$) "förklarar" x procent av den variation som ej förklaras av modell A1.
(Se vidare diskussionen av determinationsmåttets användning i kap. 6.8-6.9.)
Den multipla partiella determinationen anger hur stor andel av den tidigare oförklarade variationen som "förklaras" om en modell A1 (som exempelvis innehåller alla en- och tvåfaktorsparametrar) byggs ut till modellen A2 genom att den förra tillförs en grupp nya parametrar (vissa trefaktorparametrar). Definitionen är

$$R^2_{fA_2\cdot A_1} = \frac{\chi^2_{A1} - \chi^2_{A2}}{\chi^2_{A1}} \qquad (12.5)$$

De olika determinationsmåtten beror av täljarens och nämnarens frihetsgradtal. Liksom i regressionsanalysen kan man definiera korrigerade determinationsmått i vilka alla $\chi^2$- komponenter är dividerade med sina frihetsgradtal.

## 12.4 Politisk aktivitet genom partier

Variationerna i den politiska aktiviteten genom partier som i tidigare kapitel analyserats med hjälp av olika metoder och modeller kan också studeras med hjälp av produktmodeller (loglinjära modeller). Utgångspunkten är samma vägda frekvenstabell som i kapitel 10 vilken beräknats med hjälp av inklusionssannolikheter och svarsfrekvenser i olika kommuner. Variablerna Y (aktivitet), KV(kön), ALD (ålder), UTB (utbildning) och POL (politiskt

engagemang i familjen under uppväxttiden) ger en kontingenstabell med 2·2·3·2·2 = 48 celler varav ingen är tom.

I en produktmodell ingår alla variabler symmetriskt som faktorer. Vid tolkningen av den skattade modellen vill man ibland, såsom i detta fall, betrakta en variabel som responsvariabel och övriga som faktorer. Härav följer att man vill utröna hur aktiviteten Y beror av faktorerna men inte är intresserad av de senares inbördes relationer. Faktorernas inbördes samvariation läggs därför i denna situation alltid till grund för analysen av hur faktorerna påverkar responsvariabeln. Konsekvensen av detta är att termen KAUP (endast första bokstaven i variabelnamnen används i BMDP-programmet 4F för att specificera modell) skall ingå i modellen för den politiska aktivitetens variationer med bakgrundsfaktorerna.

## 12.5 Screening

I BMDP-programmet 4F startar sökandet efter en modell med att en följd av modeller som innehåller alla effekter med högst k faktorer skattas. Detta genomföres för

k=0: Endast referensnivån μ
k=1: Faktorerna är oberoende
k=2: Interaktion av första ordningen
k=3: - " - av andra - " -
o.s.v.

Om man fortsätter upp till k=p där p anger antalet variabler i kontingenstabellen erhålles total anpassning av modellen. Med variabeluppsättningen Y,K,A,U,P erhålles tabell 12.1.

Tabell 12.1. Utdrag ur utskrift från BMDP-programmet 4F för variabeluppsättningen Y,K,A,U och P.

***** THE RESULTS OF FITTING ALL K-FACTOR MARGINALS.
THIS IS A SIMULTANEOUS TEST THAT ALL K+1 AND HIGHER FACTOR INTERACTIONS ARE ZERO.

| K-FACTOR | D.F. | LR CHISQ | PROB. | PEARSON CHISQ | PROB. |
|---|---|---|---|---|---|
| 0-MEAN | 47 | 2530.16 | 0.0 | 2995.02 | 0.0 |
| 1 | 41 | 504.05 | 0.0 | 543.73 | 0.0 |
| 2 | 27 | 53.08 | 0.00197 | 52.91 | 0.00206 |
| 3 | 11 | 13.63 | 0.25397 | 13.98 | 0.23422 |
| 4 | 2 | 3.18 | 0.20440 | 3.15 | 0.20677 |
| 5 | 0 | 0. | 1. | 0. | 1. |

***** A SIMULTANEOUS TEST THAT ALL K-FACTOR INTERACTIONS ARE SIMULTANOUSLY ZERO.
THE CHI-SQUARES ARE DIFFERENCES IN THE ABOVE TABLE.

| K-FACTOR | D.F. | LR CHISQ | PROB. | PEARSON CHISQ | PROB. |
|---|---|---|---|---|---|
| 1 | 6 | 2026.11 | 0.0 | 2451.29 | 0.0 |
| 2 | 14 | 450.97 | 0.0 | 490.81 | 0.0 |
| 3 | 16 | 39.45 | 0.00094 | 38.94 | 0.00111 |
| 4 | 9 | 10.46 | 0.31472 | 10.83 | 0.28785 |
| 5 | 2 | 3.18 | 0.20440 | 3.15 | 0.20677 |

Övre delen av tabellen visar att modellen med samtliga trevariabeltermer inte leder till signifikant avvikelse mellan observerade och förväntade frekvenser. Nedre delen av tabellen visar att trevariabeltermerna ger en starkt signifikant förbättring av modellen (PROB=0.00094). Den slutliga modellen kommer därför med största säkerhet att innehålla ett antal trevariabeltermer. Möjligen kan det också bli aktuellt med någon term från närmast högre nivå. (Detta är den allmänna regeln. I detta fall skulle det emellertid bli en tung och svårhanterlig modell eftersom man då skulle få med fyravariabelstermer.)

Sökandet efter modell sker ofta med hjälp av test för marginell association och partiell association.
Antag att man skall konstruera en modell med 4 variabler A,B,C och D. Ett marginellt test av termen ABC jämför modellen i vilken ABC ingår som enda trevariabelterm (d.v.s. A,B,C, AB, AC, BC, ABC) med modellen utan denna trevariabelterm. Testet utgår alltså från marginaltabellen i vil-

ken man summerat över alla variabler (i detta fall D) som ej ingår i effekten ABC som skall testas.
Ett partiellt test av termen ABC jämför den fulla trevariabelmodellen A,B,C,D, AB, AC, AD, BC, BD, CD, ABC, ABD, ACD, BCD med modellen där ABC är den enda utelämnade trevariabeltermen d.v.s. A,B,C,D, AB, AC,AD,BC,BD,CD, ABD, ACD, BCD.

Som allmän regel används ofta (Brown [ 2])

- om båda testen ger signifikans bör den aktuella termen ingå i slutmodellen
- om ingetdera testet ger signifikans är termen ej aktuell för slutmodellen
- termer som är signifikanta enligt ett test men ej enligt det andra är sådana som bör studeras närmare genom jämförelser av olika par av modeller som inkluderar respektive exluderar termen ifråga.

Tabell 12.2 ger de marginella och partiella testen för variabeluppsättningen Y,K,A,U,P.

## 12.6 Modifikation för respons/faktorsituationen

Eftersom det är fråga om ett problem av "regressionstyp" med politisk aktivitet (Y) som beroende variabel och K,A,U,P som förklaringsfaktorer skall som nämnts samvariationsmönstret mellan förklaringsfaktorerna läggas som grund för analysen. Detta innebär att alla termer som kan bildas av enbart K,A,U och P automatiskt skall ingå i modellen. Termen Y skall självfallet ingå. Detta innebär endast att för givna värden på förklaringsvariablerna är sannolikheten för politisk aktivitet genom partier skild från 0.5 (grupperna politiskt aktiva och politiskt ej aktiva i populationen är inte lika stora).
Alla förklaringsfaktorer K,A,U och P torde ha klara huvudeffekter på Y eftersom KY, AY, UY och PY samtliga är starkt signifikanta enligt båda testen (PROB=0.0 eller 0.0000 i tabell 12.2).

Tabell 12.2. Politisk aktivitet genom partier. Test av partiell och marginell association med BMD-programmet P4F.

Association option selected for all terms of order less than or equal to 4

| | | Partial association | | Marginal association | |
|---|---|---|---|---|---|
| Effect | D.F. | Chisquare | Prob | Chisquare | Prob |
| K. | 1 | 60.62 | 0.0 | | |
| A. | 2 | 42.80 | 0.0000 | | |
| U. | 1 | 39.59 | 0.0 | | |
| P. | 1 | 717.74 | 0.0 | | |
| Y. | 1 | 1165.34 | 0.0 | | |
| KA. | 2 | 1.03 | 0.5984 | 2.27 | 0.3209 |
| KU. | 1 | 0.02 | 0.8825 | 0.22 | 0.6426 |
| KP. | 1 | 0.64 | 0.4223 | 2.59 | 0.1073 |
| KY. | 1 | 23.01 | 0.0000 | 26.76 | 0.0 |
| AU. | 2 | 268.03 | 0.0 | 248.65 | 0.0 |
| AP. | 2 | 5.38 | 0.0677 | 8.53 | 0.0140 |
| AY. | 2 | 45.19 | 0.0000 | 27.69 | 0.0000 |
| UP. | 1 | 25.99 | 0.0 | 37.95 | 0.0 |
| UY. | 1 | 40.80 | 0.0 | 30.90 | 0.0 |
| PY. | 1 | 45.01 | 0.0 | 58.67 | 0.0 |
| KAU. | 2 | 2.29 | 0.3175 | 2.65 | 0.2654 |
| KAP. | 2 | 1.72 | 0.4237 | 1.61 | 0.4477 |
| KAY. | 2 | 0.71 | 0.7002 | 2.13 | 0.3447 |
| KUP. | 1 | 0.71 | 0.3995 | 0.38 | 0.5385 |
| KUY. | 1 | 0.77 | 0.3788 | 1.04 | 0.3077 |
| KPY. | 1 | 3.42 | 0.0645 | 3.39 | 0.0655 |
| AUP. | 2 | 10.77 | 0.0046 | 10.99 | 0.0041 |
| AUY. | 2 | 3.69 | 0.1579 | 1.93 | 0.3808 |
| APY. | 2 | 5.55 | 0.0624 | 7.59 | 0.0225 |
| UPY. | 1 | 6.60 | 0.0102 | 8.81 | 0.0030 |
| KAUP. | 2 | 1.63 | 0.4429 | 2.15 | 0.3415 |
| KAUY. | 2 | 1.59 | 0.4507 | 2.81 | 0.2457 |
| KAPY. | 2 | 2.25 | 0.3242 | 1.85 | 0.3962 |
| KUPY. | 1 | 0.27 | 0.6045 | 0.24 | 0.6246 |
| AUPY. | 2 | 3.94 | 0.1396 | 4.22 | 0.1211 |

Då det är fråga om respons/faktorsituationer innebär trevariabeltermer i vilka Y ingår första ordningens interaktioner. KUY ger ingen signifikans enligt någondera testet. Det verkar därför rimligt att förutsätta att K och U ej samspelar i sin inverkan på Y.

Vid modellbyggnad, och särskilt i ett första sökande skede, arbetar man ofta utan strikt iakttagande av en speciell signifikansnivå. Låt säga att sannolikheter under 0.05 betraktas som klara signifikanser medan värden i intervallet

0.05-0.10 tillhör ett "grått skikt". KPY-termen ges sannolikheterna 0.0645 och 0.0655 och det får tills vidare bli en öppen fråga om termen skall ingå i modellen. Samma sak gäller termen APY som ger ett "grått värde" och en signifikans. UPY ger däremot låga sannolikheter 0.0102 och 0.0030 och får betraktas som given i modellen. Övriga trevariabeltermer innehållande Y ger höga sannolikheter och blir aktuella endast om man eljest inte finner någon modell som ger god anpassning.

## 12.7 Fortsatta modelltest

Screeningprocedurerna ovan indikerar att man kan starta med en lämplig modell, säg [KAUP, KPY, APY, UPY] och testa signifikansen hos olika termer som läggs till eller tas bort från denna modell samt testa modellernas anpassning. Det är också intressant att starta med basmodellen [KAUP, Y] som förutsätter att politiska aktiviteten Y är oberoende av faktorerna och studera hur modellens determination ökar då den byggs ut. (Se tabell 12.3.)

Tabell 12.3. Politisk aktivitet genom partier. Modelltest med hjälp av BMDP-programmet 4F.

| Modell | df | Likelihood ratio $\chi^2$ | Prob. | Multipel determ. | Partiell determ. |
|---|---|---|---|---|---|
| 1. Oberoende-modell [KAUP, Y] | 23 | 185.64 | 0.0 | (Referensmodell) | |
| 2. Modell med samtliga huvudeffekter. [KAUP,KY,AY, UY,PY] | 18 | 34.76 | 0.0101 | 0.813 | |
| Diff. vid borttagande av endast: | | | | | |
| KY | 1 | 22.82 | 0.0000 | | |
| AY | 2 | 45.42 | 0.0000 | | |
| UY | 1 | 41.20 | 0.0 | | |
| PY | 1 | 45.87 | 0.0 | | |
| 3. Modeller med samtliga huvudeffekter och en första ordningens interaktionsterm. | | | | | |
| a. Modell med KAY | 16 | 33.20 | 0.0069 | 0.821 | |
| Diff. för KAY | 2 | 1.56 | 0.4589 | | 0.045 |

forts.

3. Tabell 12.3 fortsättning

| Modell | df | Likelihood ratio $\chi^2$ | Prob. | Multipel determ. | Partiell determ. |
|---|---|---|---|---|---|
| b. Modell med KUY | 17 | 33.50 | 0.0097 | 0.820 | |
| Diff.för KUY | 1 | 1.26 | 0.2620 | | 0.036 |
| c. Modell med KPY | 17 | 30.73 | 0.0215 | 0.834 | |
| Diff.för KPY | 1 | 4.03 | 0.0447 | | 0.116 |
| d. Modell med AUY | 16 | 30.67 | 0.0148 | 0.835 | |
| Diff.för AUY | 2 | 4.09 | 0.1295 | | 0.118 |
| e. Modell med APY | 16 | 29.34 | 0.0218 | 0.842 | |
| Diff.för APY | 2 | 5.42 | 0.0664 | | 0.156 |
| f. Modell med UPY | 17 | 26.82 | 0.0607 | 0.856 | |
| Diff.för UPY | 1 | 7.94 | 0.0048 | | 0.228 |

4. Bästa modell* med alla huvudeffekter och två första ordningens interaktioner.

| | | | | | |
|---|---|---|---|---|---|
| Modell med APY, UPY | 15 | 20.90 | 0.1401 | 0.887 | |
| Diff.för APY | 2 | 5.92 | 0.0517 | | 0.222 |

5. Bästa modell* med alla huvudeffekter och tre första ordningens interaktioner.

| | | | | | |
|---|---|---|---|---|---|
| Modell med KPY,APY,UPY | 14 | 17.24 | 0.2436 | 0.907 | |
| Diff.för KPY | 2 | 3.66 | 0.0558 | | 0.175 |

6. Bästa modell* med alla huvudeffekter och fyra första ordningens interaktioner.

| | | | | | |
|---|---|---|---|---|---|
| Modell med AUY,APY,KPY,UPY | 12 | 13.14 | 0.3588 | 0.929 | |
| Diff.för AUY | 2 | 4.10 | 0.1288 | | 0.239 |

MODELL 2 med samtliga huvudeffekter ger en determination på 0.813 (eller 81.3 %).

Eftersom datorprogrammet ej tar hänsyn till det komplexa urvalsförfarandet (teori saknas) erhålles med säkerhet för låga värden i PROB-kolumnen i tabell 12.3. Modellanpassningen i termer av PROB-värdet borde därför vara bättre än den angivna medan eventuella signifikanser för enskilda termer borde vara osäkrare (högre PROB-värden än de angivna).

---

* Högsta determination

Samtliga huvudeffekter för MODELL 2 måste trots detta anses som starkt signifikanta (PROB=0.0 eller 0.0000). Modellanpassningen är så dålig (PROB=0.0101) att man trots den troliga underskattningen av sannolikheten måste underkänna MODELL 2. En eller flera interaktionstermer av första ordningen (trevariabeltermer) måste därför ingå i modellen.

Termen UPY med PROB=0.0046 införes därför vilket ökar modellens PROB till 0.0607 och determinationen till 0.856 (MODELL 3f).

Den fortsatta bedömningen är vansklig. Det förefaller rimligt såväl att stanna för MODELL 3f som att införa APY (MODELL 4) vilket ger PROB=0.1401 för modellen och PROB= 0.0517 för APY-termen med en modelldetermination på 0.887. Det är också rimligt att även inkorporera KPY vilket ger PROB=0.2436 för modellen och PROB=0.0558 för KPY-termen med en modelldetermination på 0.907.

## 12.8 Residualanalys

I tabell 12.4 ges standardiserade residualer $(O-F)/\sqrt{F}$, där O står för observerad och F för förväntad cellfrekvens. Om modellen är sann kan den standardiserade residualen approximativt betraktas som en observation från en standardiserad normalfördelning. Sannolikheten att den standardiserade residualen för en given cell skall anta ett värde utanför intervallet (-2,2) är därför approximativt 5 % om urvalet är draget som ett OSU eller som ett OSU inom strata. Med hänsyn till att urvalet är komplext torde den verkliga sannolikheten för en sådan avvikelse vara betydligt större än 5 %. De största avvikelserna i tabell 12.4 uppgår till -1.8 och 1.7 vilket därför får betraktas som högst normalt för MODELL 3f. Denna godtas därför som en adekvat modell.

Tabell 12.4. Politisk aktivitet genom partier. Standardiserade residualer för MODELL 3f. Utdrag från P4F-utskrift.

MODEL

KAUP, UPY, AY, KY.

Standardized deviates = (OBS - EXP)/SQRT(EXP) for above model

| Akt. | Pol | Utb | Åld | Kv | |
|---|---|---|---|---|---|
| | | | | Males | Females |
| Ej akt | Ej pol | Grund | Yngre | 0.2 | 0.1 |
| | | | Medel | -0.2 | -0.1 |
| | | | Äldre | -0.1 | 0.1 |
| | | Högre | Yngre | -0.1 | -0.5 |
| | | | Medel | 1.0 | 0.0 |
| | | | Äldre | -0.2 | -0.4 |
| | Pol | Grund | Yngre | -0.5 | 0.2 |
| | | | Medel | -0.6 | -0.4 |
| | | | Äldre | 0.4 | 0.9 |
| | | Högre | Yngre | 0.8 | 0.0 |
| | | | Medel | -1.3 | 0.8 |
| | | | Äldre | -0.1 | -0.2 |
| Akt | Ej pol | Grund | Yngre | -0.7 | -0.6 |
| | | | Medel | 0.5 | 0.2 |
| | | | Äldre | 0.2 | -0.2 |
| | | Högre | Yngre | 0.2 | 1.5 |
| | | | Medel | -1.8 | -0.1 |
| | | | Äldre | 0.4 | 0.8 |
| | Pol | Grund | Yngre | 1.1 | -0.5 |
| | | | Medel | 0.9 | 0.7 |
| | | | Äldre | -0.5 | -1.5 |
| | | Högre | Yngre | -1.5 | -0.0 |
| | | | Medel | 1.7 | -1.4 |
| | | | Äldre | 0.2 | 0.3 |

## 12.9 Slutlig produktmodell

Om alla termer som innehåller den beroende variabeln Y i MODELL 3f: [KAUP,UPY,KY,AY] skrivs ut erhålles

Y, KY, AY, UY, PY, UPY

Övriga termer är ointressanta eftersom de endast speglar inbördes samvariation mellan förklaringsfaktorerna. Av tabell 12.5 över parameterskattningar kan man i huvudsak avläsa vilka faktorer och faktorkombinationer som ökar respektive minskar benägenheten till politisk aktivitet genom partier samt i viss utsträckning de olika faktorernas relativa betydelse.

Däremot är det föga meningsfullt att tolka enskilda parametervärden. Då responsvariabeln är binär kan man räkna fram odds eller logitvärden, men om detta är väsentligt är det naturligtvis bättre att skatta en logitmodell direkt. Från tabell 12.5 skall vi utläsa de allmänna effekterna. Parametrarna som beskriver en viss faktors inverkan på den politiska aktiviteten (Y) får inte betraktas fristående utan endast i kombination med parametrarna för andra faktorer som den förra faktorn samspelar med. I utskriften anges alla parametrar för en viss term. Vissa är överflödiga på grund av de restriktioner som åvilar parametrarna. Det är tillräckligt att betrakta de parametrar som strukits under i tabell 12.5.

Den översta deltabellen (a) anger endast att det (i termer av geometriska medelvärden) är vanligare att väljare ej är politiskt aktiva genom partier än att de är politiskt aktiva på detta sätt.

Tabell 12.5. Politisk aktivitet genom partier. Skattningar av multiplikativa parametrar. Utdrag ur utskrift från BMDP-programmet 4F.

Estimates of the multiplicative parameters (BETA=EXP(LAMBDA))

a.

| AKT | |
|---|---|
| Ej AKT | AKT |
| 2.165 | 0.462 |

b.

| AKT | KV | |
|---|---|---|
| | Males | females |
| Ej AKT | 0.872 | 1.147 |
| Akt | 1.147 | 0.872 |

c.

| AKT | ÅLD | | |
|---|---|---|---|
| | Yngre | Medel | Äldre |
| Ej AKT | 1.318 | 0.928 | 0.818 |
| AKT | 0.759 | 1.078 | 1.223 |

d.

| AKT | UTB | |
|---|---|---|
| | Grund | Högre |
| Ej AKT | 1.177 | 0.849 |
| AKT | 0.849 | 1.177 |

e.

| AKT | POL | |
|---|---|---|
| | Ej POL | POL |
| Ej AKT | 1.265 | 0.791 |
| AKT | 0.791 | 1.265 |

f.

| AKT | POL | UTB | |
|---|---|---|---|
| | | Grund | Högre |
| Ej AKT | Ej POL | 1.092 | 0.915 |
| | POL | 0.915 | 1.092 |
| AKT | Ej POL | 0.915 | 1.092 |
| | POL | 1.092 | 0.915 |

Huvudeffekterna av bakgrundsfaktorerna på de geometriska medelfrekvenserna är följande.

Den politiska aktiviteten genom partier är

- högre för män än för kvinnor
- högre för medelålders och framför allt för äldre än för yngre personer
- högre för personer med utbildning över folkskola/grundskola än för personer med endast folkskola/grundskola
- högre för personer som vuxit upp i hem med politiskt engagemang i familjen än för personer som vuxit upp i familjer utan sådant engagemang.

Eftersom utbildning (U) och politiskt engagemang i familjen (P) samspelar måste huvudeffekterna av U och P på Y (deltabellerna d och e) modifieras med hjälp av samspelseffekten av U och P på Y (deltabell f).

- Bland personer med endast grundutbildning vilka är uppvuxna i familjer med politiskt engagemang är frekvensen politiskt aktiva högre än vad som förväntat med hänsyn till (den negativa) huvudeffekten av endast folkskola/grundskola och (den positiva) huvudeffekten av politiskt engagemang i familjen. (Med andra ord: Politiskt engagemang i familjen har haft större positiv effekt för personer med enbart folkskola/grundskola än för övriga.)

Totala effekten. För att få en bild av samspelande faktorers totala inverkan på förekomsten av politisk aktivitet måste man multiplicera ihop respektive huvudeffekter och samspelseffekter.

U och P samspelar i sin inverkan på Y. Den nedan definierade storheten $r_{ij}$ uppskattar hur stort förväntade antalet politiska aktiva inom utbildningsklass i och "politiskt engagemang"-klass j är i förhållande till vad som förväntas i denna klass med hänsyn till de övriga bakgrundsvariablerna ålder och kön. Index 2 för Y anger "aktiva".

$$r_{ij} = \theta^{UY}_{i2}\theta^{PY}_{j2}\theta^{KPY}_{ij2} \qquad (12.6)$$

För personer med enbart folkskola/grundskola och uppvuxna i hem med politiskt engagemang erhålles således

$$r_{12} = 0.849 \cdot 1.265 \cdot 1.092 = 1.17$$

Den kombinerade effekten ($r_{ij}$) av UTB- och POL-faktorerna (givet övriga faktorers värden) på förväntat antal politiskt aktiva genom partier blir för de 4 olika grupperna enligt tabell 12.6.

Tabell 12.6. Politisk aktivitet genom partier. Kombinerad effekt av utbildningsnivå och politiskt engagemang i familjen definierad som faktorn $r_{ij}$ enligt (12.6).

| | GRUND-UTB | HÖGRE UTB |
|---|---|---|
| Ej POL i familjen | 0.61 | 0.96 |
| POL i familjen | 1.17 | 1.36 |

## 12.10 Jämförelse med övriga modelltyper

"Bästa" linjära modell med samspel, "bästa" logitmodell och "bästa" produktmodell ger samtliga en adekvat beskrivning av datamaterialet.

Modellernas skilda former och de helt olika mått som ges i de olika datautskrifterna försvårar jämförelser även om dessa i och för sig låter sig göras med hjälp av olika omräkningar. De olika modelltyperna ger i stora drag samma allmänna strukturer. Skiljaktigheter beträffande förekomsten av olika samspel vad gäller logitmodell och produktmodell är huvudsakligen en följd av subjektivitet i signifikansprövningen samt av den ordningsföljd i vilken olika modellvarianter testats mot varandra.

Eftersom denna uppsats främsta syfte är att diskutera och demonstrera modellbyggnadsprocessen för de olika modelltyperna saknas anledning att göra fördjupade jämförelser mellan olika typer av modeller.

Det är emellertid uppenbart att den linjära modellen med samspel för just detta exempel ger väl så god beskrivning som de övriga modelltyperna. Då dessutom den linjära modellen med samspel både begreppsmässigt är enklare och ger enklare resultatredovisning än övriga modelltyper borde den vara ett självklart val.

# 13. SPECIELLA MODELLPROBLEM - SERVICEUTNYTTJANDE

## 13.1 Bakgrund

I kapitel 11.9 berördes Olanders och Widbergs [23] analys av hushållens utnyttjande av olika former av kommunal service. Kritiken avsåg författarnas analys som endast baserades på olika marginalfördelningar trots att problemet var multivariat.
Olander och Widberg har genomfört sin analys på endast de två fjärdedelar av respondenterna som hade högsta resp. lägsta nyttjandefrekvenserna av servicen. Tillvägagångssättet (se [23] sid. 25-32) är fullständigt olämpligt.
Det finns ett särskilt skäl att välja serviceutnyttjande som exempel i denna framställning. Här föreligger nämligen ett speciellt problem som är vanligt förekommande. Vissa celler i den kontingenstabell som beskriver den multivariata variabelfördelningen är med nödvändighet tomma. Man brukar använda uttrycket strukturella nollor för att karakterisera denna situation, vilken kräver att skattningsförfarandet för en produktmodell modifieras.

## 13.2 Biblioteksbesök

Endast respondenterna i det s.k. 5-urvalet (Kävlinge, Sjöbo, Grästorp och Luleå) tillfrågades om utnyttjandet av vissa former av kommunal service. Dessa kommuner är ej slumpmässigt valda vilket däremot gäller urvalet (OSU) av 300 personer per kommun. Resultaten kan därför endast generaliseras till befolkningen i dessa kommuner.
Varje respondent skulle bl.a. ange hur ofta någon medlem i respondentens familj besökte kommunens bibliotek.
Olander och Widberg diskuterar i rapporten [23] ett flertal bestämningsfaktorer till serviceutnyttjande inom kommunen. Vad beträffar besöksfrekvens på bibliotek anses (Olander, muntlig uppgift) att de bestämningsfaktorer som är viktiga återfinnes bland följande variabler

ålder
utbildning
förekomst av barn i förskola eller grundskola
tillgång till bil.

De data som föreligger är otillfredsställande. Ålder och utbildning är kända endast för respondenten. Olander och Widberg låter dessa egenskaper karakterisera hushållet. Sant svar på frågan "Hur ofta besöker Ert hushåll följande serviceinrättningar och affärer ?" är ett odefinierat begrepp. Svarsalternativen är

a. Aldrig.
b. Flera gånger per vecka.
c. En gång per vecka.
d. Någon gång per månad.
e. Någon gång per halvår.
f. Någon gång per år.

Även om vi förutsätter att begreppet "brukar" av respondenterna uppfattas på ett enhetligt och adekvat sätt torde det vara ganska många respondenter som inte kan ge ett korrekt besked om hushållets besöksfrekvens på biblioteket.
Dessa förhållanden medför att man drar sig för att konstruera en modell som beskriver samvariationen mellan besöksfrekvens och bakgrundsvariabler. Modellkonstruktionen nedan sker med det enda syftet att demonstrera hur man vid modellbyggnaden tar hänsyn till förekomsten av "strukturella nollor" i kontingenstabellen.

## 13.3 Variabler och klassindelning

Av 1500 utvalda respondenter har 1182 givit fullständiga svar på alla aktuella frågor. För att hålla nere antalet celler i kontingenstabellen måste antalet klasser för resp. variabel reduceras kraftigt.
Frekvensen biblioteksbesök klassificeras enligt

| | | |
|---|---|---|
| Ofta | (b, c, d ovan) | 463 personer |
| Ibland | (e, f) | 249 " |
| Sällan/aldrig | (a) | 470 " |

För bakgrundsvariablerna används klasserna

Ålder

- Unga (18-30 år)
- Medelålders (31-50 år)
- Äldre (51-70 år)

Utbildning*)

- Grund
- Vidareutbildning

Barn

- Ej skolbarn (inga barn <16 år)
- Skola (barn i skolåldern och ev. även i förskoleåldern)
- Förskola (barn i förskola men ej i skola)

Tillgång till bil

- Ej bil
- Bil

## 13.4 Kontingenstabell och modellbyggnad

En samtidig klassificering efter besöksfrekvens, ålder, utbildning, barn, biltillgång och kommun ger 3·3·2·3·2·5 = 540 celler i kontingenstabellen. Om biltillgång slopas som indelningsgrund reduceras antalet till 270 celler. Även det senare antalet är mycket högt med tanke på att antalet observationer är 1182.

I dessa tabeller förekommer många tomma celler. Vissa kombinationer av variabelvärden förekommer i populationen men ej i urvalet p.gr.a. att detta är litet i förhållande till totala antalet celler ("samplingnollor"). Vissa andra kombinationer förekommer inte i populationen. I detta fall saknas det (eller är ytterst sällsynt med) hushåll med barn i förskola eller skola och föräldrar i åldersklassen äldre. Modellen bör därför sätta förväntade antalet observationer i dessa celler till 0 (strukturella nollor).

---

*) Klassindelning enligt kap. 8.

Den specificerade produktmodellen (loglinjära modellen) förutsätts då gälla för övriga celler. Skattningsproceduren modifieras med hänsyn till den aktuella restriktionen (se Upton [27]). BMDP-programmet 4F tillåter estimation av modell med strukturella nollor.

Modell med 6 variabler

Frekvensen biblioteksbesök visar i urvalet en tydlig association med var och en av de potentiella förklaringsfaktorerna. Exempelvis ser marginalfördelningen för biblioteksbesök och tillgång till bil ut enligt tabell 13.1.

Tabell 13.1. Biblioteksbesök efter tillgång till bil.

| Tillgång till bil | Biblioteksbesök Ofta | Ibland | Sällan/aldrig | Totalt |
|---|---|---|---|---|
| Ej bil | 34 | 28 | 123 | 185 |
| Bil | 429 | 221 | 347 | 997 |
| Totalt | 463 | 249 | 470 | 1182 |

En produktmodell bör på grund av det separata personurvalet inom var och en av de fem kommunerna alltid inkludera kommuntillhörighet som en faktor. För att se vilken eller vilka potentiella förklaringsfaktorer som kunde sorteras bort skattades först en modell som innehöll alla bakgrundsvariablerna ålder (A), utbildning (U), barn (B), biltillgång (C för car) samt biblioteksbesök (L för lån) och kommuntillhörighet (K). Kontingenstabellen som erhålles vid uppdelning efter dessa 6 variabler innehåller egentligen för många celler (540 stycken vilka reduceras till 420 stycken eftersom det definieras strukturella nollor i alla celler med barn i skola eller förskola och äldre föräldrar) i förhållande till antalet observationer (1182 stycken), men på något sätt måste den eller de minst relevanta förklaringsfaktorerna identifieras.
MODELL 2 nedan (tabell 13.2) som förutsätter att det endast finns huvudeffekter av samtliga faktorer på besöksfrekvensen ger en mycket god anpassning. Ugående från en basmodell 1,

med oberoende mellan besöksfrekvens och samtliga faktorer, erhålles en relativt hög multipel determination (0.714) för MODELL 2 (se begreppsförklaringar i kap. 12).

Tabell 13.2. Modelltest, 6 variabler

| Modell | Df. | L.R.$\chi$2 | Prob. | $R^2$ | Part. $r^2$ |
|---|---|---|---|---|---|
| 1. [L,ABCUK] | 278 | 536.50 | 0.0 | - | - |
| 2. [LA,LB,LC,LU,LK,ABCUK] | 258 | 153.36 | 1.000 | 0.714 | 0.714 |

I tabell 13.3 jämförs MODELL 2 med MODELLERNA 3-6. I de senare är besöksfrekvensen oberoende av en av faktorerna A, B, C eller U. Det visar sig att MODELL 2 inte är signifikant bättre än MODELL 5 i vilken besöksfrekvens är oberoende av biltillgång (PROB$\geq$0.05).

Tabell 13.3. Modelltest, 6 visavi 5 variabler

| Modell | Df | L.R.$\chi$2 | Prob. | $\chi^2$-diff. visavi MOD.2 | Df. | Prob. för diff. |
|---|---|---|---|---|---|---|
| 2. [LA,LB,LC,LU,LK,ABCUK] | 258 | 153.36 | 1.0000 | | | |
| 3. [ ,LB,LC,LU,LK,ABCUK] | 262 | 167.85 | 1.0000 | 14.49 | 4 | 0.007 |
| 4. [LA, ,LC,LU,LK,ABCUK] | 262 | 272.50 | 0.3150 | 119.14 | 4 | 0.000 |
| 5. [LA,LB, ,LU,LK,ABCUK] | 260 | 158.34 | 1.0000 | 4.98 | 2 | >0.05 |
| 6. [LA,LB,LC, ,LK,ABCUK] | 260 | 238.61 | 0.8253 | 85.25 | 2 | 0.000 |

Testprocedurerna ovan har använts för att få en indikation på vilken faktor som i första hand kan strykas i syfte att reducera kontingenstabellen till 5 dimensioner i stället för 6. PROB-värdena i dessa tester bör uppfattas som grova approximationer.

Modell med 5 variabler

I fortsättningen analyseras kontingenstabellen LxAxBxUxK. med 270 celler av vilka 210 ej innehåller strukturella nollor.

På samma sätt som i kap. 12 startar sökandet efter en modell utifrån en tabell över partiell association och marginell

association med upp till 4-variabeltermer (ej publicerad här). Denna antyder att alla huvudeffekter (tvåvariabeltermer innehållande L) bör ingå i modellen och möjligen enstaka samspelseffekter av första ordningen (trevariabeltermer innehållande L). Dessutom ingår som alltid för asymmetriska fallet alla samspel mellan förklaringsfaktorerna inbördes.

Modellbyggnaden går till på samma sätt som i kap. 12 men redovisas här endast kortfattat i tabellform. Som basmodell användes oberoendemodellen (7) i tabell 13.4.

Tabell 13.4. Modelljämförelser baserade på kontingenstabellen LxAxBxUxK. Alla jämförelser görs relativt MODELL 8.

| Modell | | Df. | L.R.$\chi$2 | Prob. | Mult $R^2$ | Part $r^2$ |
|---|---|---|---|---|---|---|
| 7. | [L,ABUK] | 138 | 507.28 | 0.0 | | |
| 8. | [LA,LB,LU,LK,ABUK] | 120 | 102.91 | 0.8680 | 0.797 | |
| 9. | [ ,LB,LU,LK,ABUK] | 124 | 119.50 | 0.5976 | 0.764 | |
| | Diff. beroende på LA | 4 | 16.59 | <0.005 | | 0.139 |
| 10. | [LA, ,LU,LK,ABUK] | 124 | 237.74 | 0.0 | 0.531 | |
| | Diff. beroende på LB | 4 | 134.83 | 0.000 | | 0.567 |
| 11. | [LA,LB, ,LK,ABUK] | 122 | 195.14 | 0.000 | 0.615 | |
| | Diff. beroende på LU | 2 | 92.23 | 0.0 | | 0.473 |
| 12. | [LAB,LU,LK,ABUK] | 116 | 98.32 | 0.8813 | 0.806 | |
| | Diff. beroende på LAB | 4 | 2.50 | >0.60 | | 0.024 |
| 13. | [LAU,LB,LK,ABUK] | 116 | 97.64 | 0.8909 | 0.808 | |
| | Diff. beroende på LAU | 4 | 5.27 | >0.20 | | 0.051 |
| 14. | [LUK,LA,LB,ABUK] | 112 | 95.82 | 0.8628 | 0.811 | |
| | Diff. beroende på LUK | 8 | 7.09 | >0.50 | | 0.069 |

MODELL 8 med enbart huvudeffekter av förklaringsfaktorerna utgör utgångspunkt för modellarbetet. Alla huvudeffekter är starkt signifikanta (se MODELLERNA 9-11).
De möjliga interaktionseffekter på L som framkommit i tabell över marginell och partiell association testades vid jämförelse av MODELLERNA 12-14 med MODELL 8. Ingen av dessa effekter är signifikanta ens på testnivån 0.20 (se tabell 13.4). Resi-

dualmönstret för MODELL 8 är också tillfredsställande. Endast 5 av de 210 residualerna är större än 1.5. MODELL 8 kvarstår därför som slutmodell.

Slutmodell

En enligt ovan adekvat prediktionsmodell för frekvensen biblioteksbesök är således

[LA,LB,LU,LK,ABUK]

eller

$$E(f_{ijklm}) = \mu\ \theta^{L}_{i}\ \theta^{A}_{j}\ \theta^{B}_{k}\ \theta^{U}_{l}\ \theta^{K}_{m}\ \theta^{LA}_{ij}\ \theta^{LB}_{ik}\ \theta^{LU}_{il}\ \theta^{LK}_{im}$$

Envariabelparametrarna bestämmer marginalfördelningarna och är därför helt ointressanta. $\theta^{LK}$ anger att besöksfrekvensen varierar med kommun (Luleå högst och Sjöbo lägst) vilket likaså torde vara utan större intresse. Tabell 13.5 visar att följande egenskaper är förbundna med höjd besöksfrekvens

- låg ålder
- barn i skola
- vidareutbildning.

Eftersom utbildningsnivå är högre i tätorter än i glesbygd måste man emellertid fråga sig om utbildning skulle givit lika starkt utslag om det varit möjligt att låta resavstånd till bibliotek ingå i modellen. (Bibliotek ligger ju i allmänhet i tätorter även om förekomsten av bokbussar har jämnat ut tillgängligheten till biblioteksservice.)

Tabell 13.5. Huvudeffekter på besöksfrekvens på bibliotek enligt produktmodell med strukturella nollor. Utdrag ur utskrift från BMDP-programmet 4F.

ESTIMATES OF THE MULTIPLICATIVE PARAMETERS (BETA=EXP(LAMBDA)

| Ålder | Lån | | |
|---|---|---|---|
| | Ofta | Ibland | Sällan/aldrig |
| Unga | 1.016 | 1.274 | 0.773 |
| Medelåld. | 0.980 | 1.071 | 0.953 |
| Äldre | 1.005 | 0.733 | 1.357 |

| Barn | Lån | | |
|---|---|---|---|
| | Ofta | Ibland | Sällan/aldrig |
| Ej skola | 0.684 | 1.110 | 1.466 |
| Skola | 1.883 | 0.869 | 0.611 |
| Försk. | 0.864 | 1.036 | 1.117 |

| Utb. | Lån | | |
|---|---|---|---|
| | Ofta | Ibland | Sällan/aldrig |
| Grund | 0.727 | 0.890 | 1.546 |
| Vidareutb. | 1.375 | 1.124 | 0.647 |

| Lån | Kommun | | | | |
|---|---|---|---|---|---|
| | Kävlinge | Sjöbo | Grästorp | Töreboda | Luleå |
| Ofta | 0.958 | 0.819 | 1.018 | 1.107 | 1.131 |
| Ibland | 0.865 | 0.866 | 1.014 | 1.016 | 1.298 |
| Sällan/a | 1.207 | 1.411 | 0.970 | 0.889 | 0.681 |

Utslagen för de olika presumptiva förklaringsfaktorerna är förvånansvärt starka med tanke på att respondentegenskaperna ålder och utbildning fått karakterisera hushållet.
Om analysen genomförts med ett dataprogram som ej ger möjlighet att sätta strukturella nollor i vissa celler skulle resultaten blivit kraftigt snedvridna.

# R E F E R E N S L I S T A


[1 ] Bishop,Y.M.M., Fienberg, S.E. and Holland, P.W.; Discrete Multivariate Analysis: Theory and Practice. MIT Press. Cambridge MA 1975.

[2 ] Brown, M.B.; Screening Effects in Multidimensional Contingency Tables. Applied Statistics, Vol. 25, 1976 pp. 37-46.

[3 ] Draper, N. and Smith, H.; Applied Regression Analysis. 2nd ed. John Wiley & Sons New York 1981.

[4 ] Eriksson, S.; Kommunurval, väljarurval och analysansatser. Rapport 13 från kommunaldemokratiska kommittén. Ds Kn. 1982:4.

[5 ] Everitt, B.S.; The Analysis of Contingency Tables. Chapman and Hall. London 1977.

[6 ] Fienberg, S.E.; The Analysis of Cross-Classified Categorical Data. MIT Press. Cambridge MA. 1977.

[7 ] Forthofer, R.N. and Lehnen, R.G.; Public Program Analysis. A New Categorical Data Approach. Lifetime Learning Publications of Wadsworth. Belmont Ca. 1981.

[8 ] Goodman, L.A.; The Multivariate Analysis of Qualitative Data: Interactions Among Multiple Classifications. Journal of the American Statistical Association 65: 226-56. 1970.

[9 ] Goodman, L.A.; A General Model for the Analysis of Surveys. American Journal of Sociology. Vol. 77, No 6, 1972.

[10] Grizzle, J.E., Starmer, C.F. and Koch G.G.; Analysis of Categorical Data with Linear Models. Biometrics, Sept.1969, 489-504.

[11] Hanushek, E.A. and Jackson, J.E.; Statistical Methods for Social Scientists. Academic Press, New York 1977.

[12] Holt, D., Smith, T.M.F. and Winter, P.D.; Regression Analysis of Data from Complex Surveys. Journal of Royal Statistical Society A, 143, part 4, 1980 pp 474-487.

[13] Huang, D.S.; Regression and Ecomometric Methods. John Wiley & Sons. New York 1970.

[14] Intriligator, M.D.; Econometric Models, Techniques and Applications. North Holland, Amsterdam 1978. Prentice Hall, Englewood Cliffs, New Jersey 1978.

[15] Kendall, M.G. and O'Muircheartaigh, C.A.; Path Analysis and Model Building. World Fertility Survey. Technical Bulletines March 1977. No 2. International Statistical Institute. Haag.

[16] Kendall, M. and Stewart, A.; The Advanced Theory of Statistics. Vol. 2. Charles Griffin & Co Ltd. London 1979.

[17] Kleinbaum, D.G. and Kupper, L.L.; Applied Regression Analysis and Other Multivariate Methods. Wadsworth. Belmont, California 1978.

[18] Kmenta, J.; Elements of Econometrics. Mac Millan, New York 1975.

[19] Knoke, D. and Burke, P.J.; Log-Linear Models. Sage University Paper Series on Quantitative Applications in the Social Sciences, series no 07-020. Beverly Hills and London.1980.

[20] Marquart, D.W.; Generalized inverses, ridge regression, biased linear estimation and nonlinear estimation. Technometrics 12, 1970, 591-612.

[21] Mendenhall, W. and Reinmuth, J.E.; Statistics for Management and Economics. 4 th ed., Duxbury Press, Boston 1982.

[22] Nathan, G. and Holt, D.; The Effect of Survey Design on Regression Analysis. JRSS, B. Vol 42, No 3, pp 377-386. 1980.

[23] Olander, L-O. och Widberg, J.; Hushållen och serviceland-skapet. Rapport 4 från kommunaldemokratiska forskningsgruppen. Ds Kn 1981:10.

[24] Payne, C.; The Log-Linear Model for Contingency Tables. Ingår som kap. 4 i Analysis of Survey Data Vol. 2. Model Fitting. Editors O'Muircheartaigh C.A. and Payne, C. Wiley, London 1977.

[25] Snedecor, G.W. and Cochran, W.G.; Statistical Methods. 7 th ed. Iowa State University. Press.Ames, Iowa 1980.

[26] Strömberg, L. och Norell, P-O.; Kommunalförvaltningen. Rapport 15 från kommunaldemokratiska kommittén. Ds Kn 1982:8.

[27] Upton, The Analysis of Cross-Tabulated Data. Wiley & Sons. Chichester, 1978.

[28] Wallin, G., Bäck, H. och Tabor, M.; Kommunalpolitikerna. Rekrytering - arbetsförhållanden - funktioner. Rapport 8 från kommunaldemokratiska kommittén. Ds Kn 1981:17.

[29] Westerståhl, Jörgen och Johansson, Folke; Medborgarna och kommunen. Rapport 5 från kommunaldemokratiska forskningsgruppen. Departementsserien Kn 1981:12.

[30] Wonnacott, R.J. and Wonnacott, T.H.; Econometrics. 2 nd ed. John Wiley and Sons. New York 1979.

[31] Younger, M.S.; A Handbook for linear regression. Wadsworth. Belmont. California 1979.

RESEARCH REPORT

1983:1 Flood, L.: Time allocation to market and non-market activities in Swedish households.

1983:2 Eriksson, S.: Analys av kategoriska data. En metodstudie i anslutning till statsvetenskaplig forskning.