STATISTISKA CENTRALBYRÅN 1980-10-22

PROMEMORIOR FRÅN P/STM
NR 4

HANDLEDNING I AID-ANALYS

AV
ANDERS NORBERG

INLEDNING

TILL

Promemorior från P/STM / Statistiska centralbyrån. – Stockholm : Statistiska centralbyrån, 1978-1986. – Nr 1-24.

Efterföljare:

Promemorior från U/STM / Statistiska centralbyrån. – Stockholm : Statistiska centralbyrån, 1986. – Nr 25-28.

R & D report : research, methods, development, U/STM / Statistics Sweden. – Stockholm : Statistiska centralbyrån, 1987. – Nr 29-41.

R & D report : research, methods, development / Statistics Sweden. – Stockholm : Statistiska centralbyrån, 1988-2004. – Nr. 1988:1-2004:2.

Research and development : methodology reports from Statistics Sweden. – Stockholm : Statistiska centralbyrån. – 2006-. – Nr 2006:1-.

Promemorior från P/STM 1980:4. Handledning i AID-analys / Anders Norberg.
Digitaliserad av Statistiska centralbyrån (SCB) 2016.

urn:nbn:se:scb-PM-PSTM-1980-4

STATISTISKA CENTRALBYRÅN 1980-10-22

PROMEMORIOR FRÅN P/STM
NR 4

HANDLEDNING I AID-ANALYS

AV
ANDERS NORBERG

INNEHÅLLSFÖRTECKNING TILL HANDLEDNING I AID-ANALYS

Handledning i AID-analys

1 Inledning

Syftet med denna handledning är att beskriva AID-metoden för den som inte har några speciella förkunskaper i statistisk analys och att ge vissa tips om när det är lämpligt resp olämpligt att använda metoden. Vissa avsnitt innehåller några formler vilket kräver viss vana vid statistisk terminologi, men man kan förstå vad metoden innebär utan att man fördjupar sig i dessa avsnitt.

AID är i första hand en metod för s k dataanalys. Detta betyder att AID är ett hjälpmedel för att ganska förutsättningslöst studera sambandsstrukturen hos variablerna i en insamlad datamängd. En klassisk metod som regressionsanalys kan användas för dataanalys men är inte lika flexibel. Regressionsanalys har å andra sidan förutsättningar för att användas i en strikt statistisk analys med odiskutabla inferensmöjligheter. Orsaken till dessa skillnader mellan metoderna ligger i att regressionsanalysen bygger på en matematiskt statistisk modell i vilken vi skattar parametrar. Via denna modell kan man dra slutsatser om den studerade företeelsens egenskaper. Dataanalys ger förhoppningsvis uppslag till en rad hypoteser vilka med ett nytt urval kan testas med t ex regressionsanalys. Dataanalysen ger inte underlag för att dra några slutsatser med större tillförlitlighet.

För SCBs del kan AID-metoden vara ett värdefullt redskap för t ex stratifiering vid urvalsplanering. I Hushållsbudgetundersökningen har den använts för ett liknande ändamål.

Om Du fortsätter att läsa denna handledning så finner Du nog att det är lättare att förstå hur metoden fungerar än de filosofiska tankegångarna kring hur resultaten kan tolkas och användas. Försumma emellertid inte detta senare område.

2 Beskrivning av AID-metoden

AID är en förkortning av Automatic Interaction Detector vilket alltså antyder att denna metod skall kunna upptäcka samspel (interaktioner). Metodens likhet med regressionsanalys är att man vill förklara en beroende variabels variationer med hjälp av ett antal oberoende variabler.

AID introducerades 1964 av Morgan och Sonqvist vid the University of Michigan, vilka också programmerade metoden för dator. Programversionen AID3, som finns vid SCB, ingår i programpaketet OSIRIS och beskrivs i Sonqvist et at (1973). AID har som analysmetod utvecklats parallellt med programmet. Detta gäller även för många andra dataanalysmetoder som t ex klusteranalys. Det är bara i mycket enkla tillämpningar som man skulle klara av beräkningarna utan datorkapacitet.

Beskrivningen av AID-metoden i denna handledning är ofta lika mycket en beskrivning av AID-programmet. Programmets begränsningar och möjligheter definierar i vissa hänseenden metodens egenskaper.

Innan AID-metoden och -programmet ytterligare beskrivs skall vi berätta vad vi menar med beroende och oberoende variabler samt samspelseffekter mellan de oberoende variablerna.

2.1 Begreppen "beroende" och "samspel"

Låt oss ta följande uttalanden som illustrationsexempel:

- Att sjöarna i Sverige försuras beror bl a på luftföroreningar från industrierna i Västtyskland.

- Anställdas löner beror på yrke och ålder.

- Sjukfrånvaron hos barn på daghem är högre för yngre än för äldre barn.

Ordet bero är i svenska språket synonymt med betingas av, bestämmas av etc, dvs det används när man talar om relationer med orsak och verkan. På detta sätt används ordet i hypotesen om sjöarnas försurning ovan. En eller flera sådana hypoteser som berör samma problemområde utgör en förenklad beskrivning av en del av den komplexa verkligheten. Den statistiska modellen är översättning av den förenklade beskrivningen till matematiska uttryck bestående av stokastiska variabler och okända koefficienter.

Ett exempel på statistisk modell är den linjära regressionsmodellen med en beroende variabel Y och en eller flera oberoende variabler X_i. I en sådan modell skulle försurningsgraden vara den beroende variabeln och föroreningsgraden en oberoende variabel. Försurningen antas vara proportionell mot föroreningsgraden. Modellens parameter är proportionsfaktorn som vi skattar med data från experiment eller observationer i naturen.

Med den linjära regressionsmodellen skulle vi också kunna studera sambandet mellan barnens sjukfrånvaro som beroende och ålder som oberoende variabel utan att sambandet är direkt kausalt; det är ju inte åldern som utgör motståndet mot infektioner utan egenskaper som barnet förvärvar under sin levnad. Observera att det inte stod att sjukfrånvaron beror av ålder. Ordet beroende används i den statistiska modellen utan att kausalitet nödvändigtvis föreligger i de förhållanden som studeras.

AID bygger inte på någon matematisk modell som regressionsanalys. Variabeluppsättningen är emellertid av samma slag med en beroende och ett antal oberoende variabler och de samband som analyseras kan vara av kausal natur men det är inte nödvändigt.

Samspel[1] kan finnas mellan två oberoende variabler X_1 och X_2 i dessas samband med beroende variabeln Y. Att samspel

---

[1] Samspel = interaktion

saknas betyder att effekten av X_1 på Y är oberoende av värdet på X_2 och vice versa. När samspel finns så är skillnaden mellan Y-värdena för två olika X_1-värden också beroende av värdet på X_2. Anställdas löner är högre för akademiker än för ickeakademiker och lönerna är högre för äldre än för yngre. Detta är två fristående beskrivningar av sambanden mellan inkomst å ena sidan och yrke och ålder å andra sidan där samspel mellan yrke och ålder inte beaktas. Löneskillnader mellan olika åldersgrupper är emellertid mycket kraftigare för akademiker än för ickeakademiker, vilket betyder att samspel finns mellan yrke och ålder. Samspel kan även finnas mellan tre eller flera oberoende variabler.

Med AID-analys undersöker man strukturen i en insamlad datamängd för att få uppslag till hypoteser. Av namnet att döma borde AID vara speciellt lämpad för att upptäcka interaktioner. Detta är emellertid inte helt säkert, speciellt inte om man nöjer sig med att bara studera det erhållna träddiagrammet. Läs mer om detta i avsnitt 5.1.

2.2 Klyvning av data

AID är en stegvis delningsteknik. En hierarki av grupper byggs upp steg för steg. I varje steg sker efter en viss räkneprocedur en klyvning av en grupp i datamaterialet. När tekniken i första steget är klargjord är det lätt att förstå även resten.

En given datamängd beskrivs med bl a antalet objekt, medelvärdet på den variabel som man studerar (beroende variabeln) och denna variabels variation. Antag att vi också har samlat information om ett antal bakgrundsfaktorer som kan tänkas ha samband med den beroende variabeln. Denna information representeras av ett antal oberoende variabler. Det allra enklaste sättet att visa på ett samband mellan den beroende variabeln och en oberoende variabel är att med hjälp av den oberoende variabeln dela materialet i två grupper i vilka den beroende variabelns medelvärden är olika.

Låt oss ta barnens sjukfrånvaro som exempel. I en urvalsundersökning har vi funnit att sjukfrånvaron i procent av överenskommen närvarotid för alla barnen i genomsnitt är 10.4 %. Vi tror att sjukfrånvaron är högre för de små barnen och delar hela materialet i barn under 3 år och barn över 3 år. Dessa grupper har den genomsnittliga sjukfrånvaron 15.9 % resp 8.5 % dvs det finns en klar skillnad. Denna analys kan illustreras i t ex en tabell:

Åldersgrupp	Genomsnittlig sjukfrånvaro
Samtliga	10.4 %
därav under 3 år	15.9 %
över 3 år	8.5 %

2.2.1 Kvadratsummans uppdelning

Detta avsnitt innehåller några enkla formler som kan vara bra att förstå när avsnitt 2.3 skall läsas.

Låt oss kalla den beroende variabeln Y och de observationer vi har på denna variabel $y_1, y_2, \ldots, y_n$. Medelvärdet av alla observationer är $\bar{y}$. Totala summan av de kvadrerade avvikelserna mellan observationerna $y_1, y_2, \ldots y_n$ och $\bar{y}$ kallar vi TSS (Total Sum of Squares)

$$TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

Dela datamaterialet med hjälp av den oberoende variabeln X_1 i två grupper A och B med n_A resp n_B observationer så att X_1 endast antar vissa värden i grupp A och resterande värden i grupp B. Medelvärdena är $\bar{y}_A$ och $\bar{y}_B$. Vardera gruppen har nu en egen TSS som definieras på samma sätt som ovan. Summan av dessa två kvadratsummor, TSS_A resp TSS_B, kallar vi WSS (Within groups Sum of Squares) till skillnad från BSS (Between groups Sum of Squares)

$$WSS = \sum_{i=1}^{n_A} (y_i - \bar{y}_A)^2 + \sum_{i=1}^{n_B} (y_i - \bar{y}_B)^2$$

$$BSS = n_A \cdot (\bar{y}_A - \bar{y})^2 + n_B \cdot (\bar{y}_B - \bar{y})^2$$

Det är känt och kan lätt visas att

$$TSS = WSS + BSS$$

Låt Y vara sjukfrånvaro i procent för barn på daghem. Låt grupp A vara barn under 3 år och grupp B vara barn över 3 år. Medelvärdena är $\bar{y}_A = 15.9$ och $\bar{y}_B = 8.5$ dvs det finns en skillnad. Om det inte varit någon skillnad skulle BSS vara noll och TSS skulle vara lika med WSS. Nu är $BSS = 7.6 \cdot 10^5$ jämfört med $TSS = 115 \cdot 10^5$. Ett siffermått på styrkan i sambandet mellan variablerna sjukfrånvaro och ålder är $BSS / TSS = 0.066$.

Medelvärdet $\bar{y}$ är för varje enskilt objekt en mycket grov approximation av y_i och därför är den totala summan av de kvadrerade avvikelserna mellan approximationen och de observerade värdena stor (TSS). Medelvärdena $\bar{y}_A$ och $\bar{y}_B$ är för objekten i respektive grupp bättre approximationer, varför summan av de kvadrerade avvikelserna mellan y_i och approximationerna $\bar{y}_A$ resp $\bar{y}_B$, WSS, är mindre än TSS. Den relativa minskningen av kvadratsumman är BSS/TSS. Det ligger nära till hands att kalla denna kvot förklaringsgrad och uttrycka den i procent. Det är så man gör i AID-analys och regressionsanalys med dummyvariabler där materialet delas efter någon variabel i två grupper.

I linjär regressionsanalys med en kontinuerlig oberoende variabel ersätter vi WSS med $SS_{residual}$ och BSS med $SS_{regression}$.

$$SS_{residual} = \sum_{i=1}^{n} (y_i - \hat{\alpha} - \hat{\beta}x_i)^2$$

$$SS_{regression} = \sum^{n} (\hat{\alpha} + \hat{\beta}x_i - \bar{y})^2$$

där $\hat{\alpha}$ och $\hat{\beta}$ är skattningar av regressionskoefficienterna. På motsvarande sätt kallar vi $SS_{regression}$ /TSS förklaringsgrad och menar att vi lyckats förklara en del av variationerna hos variabeln Y med hjälp av en variabel X. Observera att vi inte därmed kan påstå att variabeln X orsakar att Y varierar.

2.3 AID-metoden

När man har flera oberoende variabler[1] vilka var och en antar ett flertal värden så skulle en successiv dataklyvning 'för hand' bli långt ifrån optimal. Med optimal menar vi att förklaringsgraden (BSS/TSS) skall vara så stor som möjligt. Det blir helt enkelt för stort räknearbete för att alla delningsmöjligheter skall kunna beaktas. AID-programmet väljer i varje steg den bästa klyvningen och med steg menas då att en delgrupp skall delas ytterligare en gång i två grupper.

Sökningen efter den bästa klyvningen har två dimensioner:

1° för varje oberoende variabel skall den bästa grupperingen av variabelvärden sökas

2° den oberoende variabel som åstadkommer den bästa delningen skall väljas.

Denna tvåstegs-sökning ställer kravet på de oberoende variablerna att de bara antar ett begränsat antal värden. Programmet tillåter värden mellan 0 och 31 men man bör inte gärna ha fler olika värden än 10. Om en variabel av naturen skulle vara kontinuerlig får man i förväg göra en klassindelning. Efter första delningen av datamaterialet finns alltså två grupper som kan delas enligt samma metod. Programmet väljer att först försöka dela den grupp som har den största andelen av de oförklarade variationerna eftersom man kan förvänta sig att göra den största "förklaringen" i den. Det är dock inte alls uteslutet att man kan få stora variansförklaringar i grupper som delas sent i analysen. För att en delning skall kunna ske i programmet ställs tre villkor:

i totala antalet delningar skall vara mindre än vad som angetts i förväg

ii varje grupp måste ha fler objekt än ett i förväg specificerat antal, (dvs minsta gruppstorlek)

---

[1] I litteraturen används ordet prediktor synonymt med oberoende variabel i samband med AID-analys.

iii Kriteriefunktionen vid delning av grupp L är

$$\lambda_L = BSS_{Li,\ Lj} / TSS$$

där $BSS_{Li,Lj}$ är mellangruppskvadratsumman vid delning av grupp L i grupperna Li och Lj. Då TSS i hela analysen har samma värde kan det tyckas onödigt att dividera med TSS, men det ger λ_L innebörden av förklaringsbidrag från delningen av grupp L. Delning sker endast om $\lambda_L \geq \lambda_o$ där λ_o angetts i förväg. Flera forskare har funderat på hur man skall sätta ett lämpligt värde på λ_o men det sannolikhetsteoretiska resonemanget blir synnerligen komplicerat. Se kapitel 3.

Vi skiljer på två typer av oberoende variabler:

1 En oberoende variabel deklareras som monoton om den har följande egenskaper:
- den skall vara mätt på åtminstone ordinalskala, dvs den har en naturlig ordning i variabelvärdena. Exempel: Ålder indelad i fem klasser
- delning skall ske så att den ena gruppen har värden under en viss nivå på variabeln och den andra gruppen har värden över nivån.

Exempelvis är ålder en monoton variabel i sambandet med daghemsbarnens sjukfrånvaro men skulle inte vara det i sambandet med svenskarnas inkomster, där medelålders personer har större inkomster än unga och pensionärer. En monoton variabel med K klasser ger K-1 olika delningar av datamaterialet att välja bland.

2 Motsatsen till monoton är fri (fri från restriktioner), vilket gäller för nominalskalevariabler hos vilka variabelvärdena saknar ordning, t ex alternativen promenad, bil, buss eller tåg för att ta sig till arbetet. Variabler med ordning där ordningen inte stämmer överens med beroende variabelns ordning deklareras också som fria. En fri variabel med K klasser ger $2^{K-1}-1$ olika delningsmöjligheter. En fri variabel ger alltså upphov till många fler delningsmöjligheter än en monoton när K blir stort.

2.4 Ett exempel: Daghemsundersökningen

SCB gjorde 1977 en stor undersökning av daghemmen (se SM S 1979:1). En viktig frågeställning i undersökningen var vilka orsaker som fanns till barnens sjukfrånvaro. Detta har belysts med tabeller i SMet och med regressionsanalys i bilagan till detsamma. Man kan emellertid ifrågasätta regressionsmodellen därför att den förutsätter linjära och additiva effekter av ålder, bostadsmiljö och personal på barnens sjukfrånvaro.

I detta avsnitt skall vi som illustration göra en AID-analys på samma data men med ganska få variabler och följa analysen steg för steg.

Analysen görs av de 5 953 poster som ej har något partiellt bortfall. Varje objekt (barn) vägs med uppräkningsfaktorerna i undersökningen för att kompensera för att vissa grupper av barn är överrepresenterade i urvalet jämfört med verkligheten. Detta innabär att varje term i kvadratsummorna TSS, BSS och WSS multipliceras med uppräkningsfaktorn för motsvarande urvalsobjekt. I övrigt är analysen identisk med vad som redan beskrivits. Angående vägning, se avsnitt 4.1.

Beroende variabel Y är sjukfrånvaro i procent av överenskommen närvaro. Oberoende variablerna och dessas kategorier framgår av följande tablå. M betyder monoton och F fri variabel.

Variabel	Kategorier
Ålder (M)	1, 2, 3, ..., 9 år
Bostad (F)	1 = villa/radhus/jordbruksfastighet
	2 = lägenhet i förort/ytterområde
	3 = lägenhet centralt i storstadsområde eller tätort
Färdsätt (F)	1 = promenad
	2 = bil
	3 = tåg/buss
	4, 5, 6 = annat, t ex kombinationer av 1-3
Region (F)	1 = storstadsregion
	2 = övriga landet
Personaltäthet (F)	0-1.9, 2.0-3.9, 4.0-5.9,... barn/person-årsverk

Vi sätter följande villkor på analysen:

i Högst 20 delningar
ii Minsta gruppstorlek = 25 barn
iii Minst 0.3 % av hela materialets TSS måste förklaras i varje delning, dvs $\lambda_o = 0.003$.

Steg 1

Medelvärdet på Y-variabeln är i hela materialet som vi kallar grupp 1, 10.4. Den totala kvadratsumman är $11.5 \cdot 10^6$. Oberoende variabel nr 1 är Ålder. Det är en monoton variabel som kan dela materialet på följande 8 olika sätt eftersom den har 9 klasser.

Ålder i grupp A	Ålder i grupp B	100 · λ
1	2-9	2.6
1-2	3-9	5.4
1-3	4-9	6.6
1-4	5-9	5.0
1-5	6-9	3.9
1-6	7-9	1.3
1-7	8-9	< 0.3
1-8	9	< 0.3

Vi ser att med variabeln Ålder som prediktor erhålls bästa delningen med klasserna 1-3 år resp 4-9 med $\lambda = 0.066$.

Med variabeln bostad som prediktor får vi tre olika grupperingar.

Bostadskod i grupp A	Bostadskod i grupp B	100 · λ
kod 1	kod 2,3	1.4
kod 2	kod 1,3	1.2
kod 3	kod 1,2	?

Programmet beräknar inte λ-värden för alla tänkbara delningar med fria variabler. Först beräknas i stället den beroende variabelns medelvärde för varje kategori hos den oberoende variabeln. Därefter sorteras kategorivärdena i ordning efter medelvärdena varefter sökningen efter bästa λ-värde blir lika enkel som för monotona variabler.

För variabeln Bostad analyserar programmet alltså bara två delningar av de tre möjliga. Vi har följande medelvärden:

Bostad	Proc sjukfrånvaro
kod 1	8.7
kod 2	12.2
kod 3	10.8

Det är meningslöst att slå samman koderna 1 och 2 i hopp om att finna en skillnad mot kod 3 som är större än skillnaden när koderna 1 och 3 eller 3 och 2 slås samman.

På samma sätt som för Bostad görs analys även för de tre fria variablerna Färdsätt, Region och Personaltäthet.

Andra momentet i valet av bästa dataklyvning är att välja den bästa av de bästa klyvningarna. Vi jämför alltså variablernas λ_{max}-värden

Variabel	$100 \cdot \lambda_{max}$
Ålder	6.6
Bostad	1.4
Färdsätt	0.8
Region	1.8
Personaltäthet	2.6

Den allra bästa klyvningen görs alltså med variabeln Ålder i klasserna 1-3 år och 4-9 år och λ_{max} är större än gränsvärdet 0.3 % varför delning accepteras. De två grupperna numreras för ordningens skull efter medelvärdet på beroende variabeln.

Nya grupper	Antal urvals-objekt	Antal uppvägda objekt	Medel-värde	TSS
Grupp 2 (4-9 år)	3 429	54 567	8.5	$6.5 \cdot 10^6$
Grupp 3 (1-3 år)	2 524	18 814	15.9	$4.2 \cdot 10^6$

Följ analysens steg även i diagram 1 på sidan 16.

Steg 2

Grupp 2 är nu den odelade grupp som har störst inomkvadratsumma TSS_L. Därför försöker vi dela den.

Variabeln Ålder kan fortfarande vara kandidat som delande variabel vid delning av grupp 2 men med ett mindre värdeförråd. I övrigt följer vi samma algoritm som i steg 1. Jämförelsen mellan variablernas bästa delningar blir

Variabel	$100 \cdot \lambda_{max}$
Ålder	0.8
Bostad	0.5
Färdsätt	0.2
Region	0.7
Personaltäthet	0.6

Det är nu relativt små λ_{max}-värden men vi accepterar en delning med åldern. Observera att Ålder inte delar så mycket bättre än Region. Följande grupper kan nu delas

Grupp	Antal urvals-objekt	Antal uppvägda objekt	Medel värde	TSS_L
3 (1-3 år)	2 524	18 814	15.9	$4.2 \cdot 10^6$
4 (6-9 år)	1 615	27 462	7.2	$2.4 \cdot 10^6$
5 (4-5 år)	1 814	27 105	9.9	$3.9 \cdot 10^6$

Steg 3

Vi försöker nu dela grupp 3 eftersom den har störst TSS_L. Resultatet liknar steg 2 med Ålder som "bästa" delare, men det skiljer väldigt lite mellan Ålder och Färdsätt. Vi skulle kanske få en analys med minst lika hög förklaringsgrad om vi valde Färdsätt i stället för Ålder i detta steg.

Variabel	$100 \cdot \lambda_{max}$
Ålder	0.89
Bostad	0.67
Färdsätt	0.85
Region	0.75
Personaltäthet	0.40

Nu finns fyra kandiderande grupper:

Grupp	Antal urvals-objekt	Antal uppvägda objekt	Medel-värde	TSS
4 (6-9 år)	1 615	27 462	7.2	$2.4 \cdot 10^6$
5 (4-5 år)	1 814	27 105	9.9	$3.9 \cdot 10^6$
6 (2-3 år)	2 243	16 909	15.1	$3.5 \cdot 10^6$
7 (1 år)	281	1 905	22.8	$0.6 \cdot 10^6$

Steg 4

4-5 åringarna i grupp nr 5 har den största inomvariationen m a p sjukfrånvaron. De oberoende variablerna har följande mått på samvariationen med beroende variabeln:

Variabel	$100 \cdot \lambda_{max}$
Ålder	0.05
Bostad	0.47
Färdsätt	0.17
Region	0.80
Personaltäthet	0.49

Äntligen får vi se ett exempel på en annan variabel än Ålder som delare. Bland 4-5 åringarna finns i datamaterialet en medelvärdesskillnad på 3.7 procentenheter i sjukfrånvaro mellan barn i storstad och barn i övriga landet.

Steg 5-7

I steg nummer fem delas 2-3 åringar upp efter hur de kommer till daghemmet. 4-5 åringarna i storstad delas i steg nummer sex efter variabeln personaltäthet i två, vid första anblicken, svårförklarade grupper. Ytterligheterna 0-1.9 resp över 10 barn per personårsverk är emellertid relativt små grupper. Tolkningen blir att sjukfrånvaron är lägre när det är fler än sex barn per personårsverk än när det är färre än sex barn. Motsvarande resultat fick man med regressionsanalysen.

Steg nummer sju är det sista som uppfyller villkoren för delning. Det är 1-åringarnas sjukfrånvaro som är olika i storstad och övriga landet.

Vi kan sammanfatta denna analys med AID-trädet på nästa sida. Observera att när en AID-analys endast presenteras med ett sådant träddiagram så har en stor del av analysens resultat undanhållits läsaren.

Analysen har gett uppslag till följande hypoteser (om de inte var tämligen givna på förhand):

- barns benägenhet att vara sjuka "beror" i första hand på barnets ålder; yngre barn har större sjukfrånvaro än äldre

- sjukfrånvaron är högre i storstad än i övriga landet. Skillnaderna är större för yngre barn än för äldre. (det finns en interaktion mellan Ålder och Region)

- 2-3 åringars sjukfrånvaro är lägre för barn som får åka bil till daghemmet än för dem som går eller åker kommunalt. Färdsättet har större betydelse än var man bor. (Färdsätt, Bostad

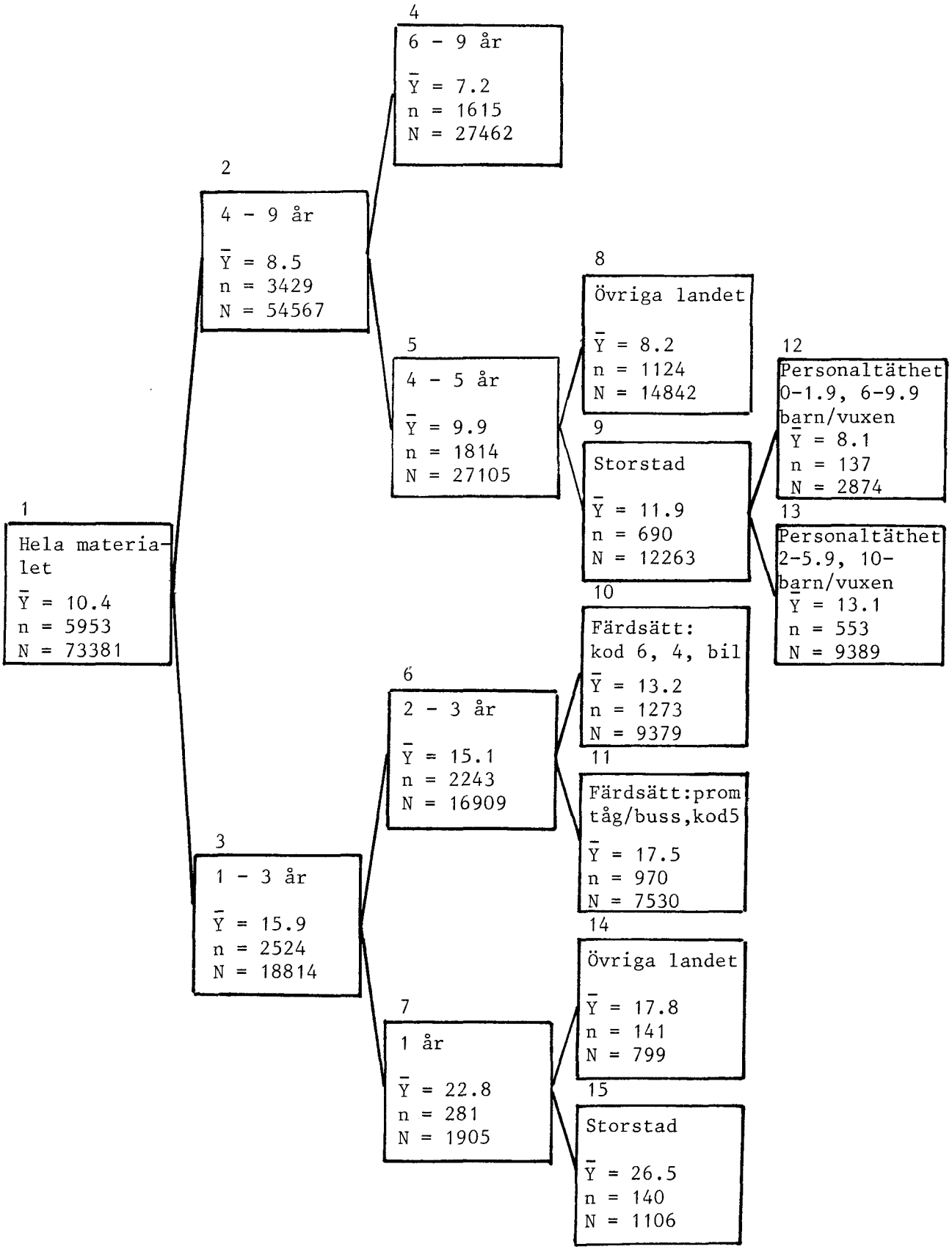
och Region är emellertid nästan lika bra delare i denna grupp. Hypotesen om att just färdsättet är viktigast är mycket vag.)

- sambandet med personaltätheten är tvärtemot vad man kanske har gissat. Bland 4-5 åringar i storstad är sjukfrånvaron större när det finns 2-6 barn/personal än när det är 6-10 barn/personal. Om det är barnens höga sjukfrånvaro som kräver mer personal eller om den höga personaltätheten orsakar högre sjukfrånvaro säger AID-metoden inget om. Det kan givetvis finnas andra förklaringar också.

Den som använder AID-metoden för att analysera datamaterialet och sedan ämnar redovisa undersökningen i traditionell tabellform finner att all redovisning av sjukfrånvaron mot bakgrundsfaktorer måste även ha en åldersuppdelning. Analysen ger ett förslag med fyra åldersklasser. Det är dock inte säkert denna uppdelning i fyra klasser som är den bästa.

Analysen har en tämligen låg förklaringsgrad. Varje steg har, med undantag av första, förklaringsgrader under 1 %. Det går emellertid inte att generellt säga vad som är låg eller hög förklaringsgrad, det beror på tillämpningen ifråga.

Diagram AID-analys av sjukfrånvaro på daghem vägt med inverterade urvalssannolikheter



3 Statistisk inferens vid AID-analys

Det finns en del skrivet om inferens vid AID, Avén (1974), Kass (1975) och Scott & Knott (1976), men området är inte färdigutforskat. Låt oss till en början reda ut begreppet inferens.

3.1 Allmänt om statistisk inferens

Att göra statistisk inferens är att med hjälp av statistiska metoder dra slutsatser om en mängd objekt. Slutsatsen kan t ex avse summan av en variabels värden i en ändlig population eller form och styrka i ett kausalt samband mellan två eller flera variabler. Vid slutsatser av det senare slaget bortser man ofta från om populationen är ändlig[1] eller om objekten är resultat av en serie experiment. I samhällsstatistiska undersökningar är objektmängderna ändliga.

Att från ett urval göra inferens om den ändliga populationen, avseende totalen eller medelvärdet på någon variabel, innebär att med viss, känd eller skattningsbar, osäkerhet dra slutsatser om populationen av samma slag som vi skulle göra om urvalet var identiskt med hela populationen.

Slutsatser om samband mellan olika fenomen hos en viss typ av objekt baseras på skattningar av parametrar i en sambandsmodell. Objektens variabelvärden betraktas härvid som realiseringar av stokastiska variabler för vilka vissa sannolikhetsmodeller gäller. Att populationen faktiskt är ändlig innebär begränsningen att urvalet av objekt inte kan bli större än populationen (exempelvis kan analys av Sveriges kommuner baseras på högst 277 objekt). En annan begränsning är att man i allmänhet inte kan mäta effekter av olika politiska åtgärder m m för att göra uttalanden om klausaliteten i samband. Av daghemsundersökningen skulle vi inte kunna dra slutsatser om vilken inverkan personaltätheten har på barnens sjukfrånvaro, bara att det finns samband mellan dessa variabler. För ett fylligare resonemang rekommenderas Wretman (1980).

---

[1] Man säger ibland att den ändliga populationen är ett urval från en superpopulation.

3.2 Signifikansproblemet

En rad forskare, bl a Avén (1974), har studerat det s k signifikansproblemet vid AID-analys vilket innebär att man försöker avgöra om en delning motsvarar en signifikant skillnad mellan gruppmedelvärdena med beaktande av alla tänkbara delningsmöjligheter. Även om delningarna är signifikanta skulle andra AID-träd med delvis några av de andra variablerna i analysen eller andra kategorigrupperingar också kunna vara "signifikanta". Att ett AID-träd är "signifikant" innebär alltså inte att vi med någon större säkerhet skulle få exakt samma AID-träd med annat urval. Det är emellertid väldigt bra att kunna avgöra om en delning troligtvis motsvarar en verklig skillnad mellan grupperna. Problemet med denna signifikansanalys är att den blir tämligen arbetskrävande men samtidigt får man ännu mer insikt i datamönstret. En beskrivning av signifikansproblemet ges i Norberg (1980).

Sammanfattningsvis är AID ämnat för dataanalys, dvs när vi inte gör någon verklig statistisk inferens utan undersöker datamängden för att få uppslag till modeller. Detta innebär att man i viss mån tänker i modelltermer men analysen ger inga slutsatser som vi vill presentera på annat sätt än som hypoteser.

4 Analys av surveydata

I detta avsnitt skall vi ta upp några problem som gäller all analys av data insamlade i surveyer till skillnad från experimentella data.

4.1 Urvalsdesign

I våra urvalsundersökningar vid SCB har vi mycket sällan ett obundet slumpmässigt urval. Vanligast är att man stratifierar populationen och gör slumpmässiga urval inom strata vilket alltså ger objekten lika urvalssannolikheter stratumvis men olika mellan strata. Vid PPS-urval, vilka inte är så vanliga i individ- och hushållsundersökningar, får samtliga objekt olika urvalssannolikheter.

När man estimerar antalet objekt i en tabellcell eller medelvärdet på någon variabel är det självklart att man väger urvalsobjekten med inverterade urvalssannolikheter. Detta motsvarar en inferens till den faktiska ändliga populationen enligt vårt resonemang i avsnitt 3.1. Om man av någon anledning vill göra en tabell som endast avser urvalet skall man, lika självklart, inte väga objekten med olika vikter. Inferens via den statistiska modell som antas ha genererat populationen kan göras enligt två alternativ (Wahlström & Lindström 1971):

i direkt från urvalet via modellen vilket betyder att hänsyn inte tas till urvalsdesignen

ii från urvalet via populationen och därefter via modellen vilket innebär vägning.

Återigen skall det påpekas att AID-metoden saknar modell och att ovanstående allmänna inferens- och estimationsresonemang inte gäller. Paralleller kan dock dras.

Vägning vid AID-analys innebär att varje term i kvadratsummorna multipliceras med objektens vikter, se avsnitt 2.2.1. AID-metoden tar i hög grad hänsyn till de uppvägda antalen objekt i grupperna. En delning görs nämligen sällan i en grupp med mycket få och en grupp med många objekt. Det har därför betydelse vilka vikter som väljs eftersom en liten grupp i urvalet kan bli större efter vägning.

En ren dataanalys av variabelsamband utan stark koppling till den faktiska ändliga populationen kan göras med alla vikter lika med ett. Å andra sidan kan urvalet ha överrepresenterat en viss grupp av objekt på ett sätt som vi inte gillar eftersom förhållandena i den gruppen då dominerar i resultaten. Vägning med inverterade urvalssannolikheter kan lösa detta problem. Det kan också vara så att varken urval eller population är så fördelad på bakgrundsvariabler som vi önskar. Då kan man mycket väl konstruera andra vikter så att man får önskad fördelning på bakgrundsvariablerna. (Paralleller

kan dras med i och ii ovan som säger att inferens via modell kan baseras på ovägda eller vägda parameterskattningar.) Det skall dock påpekas att för att Avéns signifikansanalys skall fungera tillfredsställande bör alla vikter vara lika.

När AID-metoden används för andra speciella ändamål bör man välja vikter med hänsyn till syftet. Vid analyser som syftar till att ge information om den faktiska ändliga populationer, t ex för val av strata, bör man således väga med inverterade urvalssannolikheter.

4.2 Bortfall och mätfel

Bortfallsfel och mätfel snedvrider givetvis resultaten vid AID-analyser på liknande sätt som vid användning av andra analysmetoder. Bortfallsbehandlingen vid analys med AID kan i stort sett gå till på samma sätt som vid andra analysmetoder. Objektsbortfall ingår ej i datamaterialet och kan lämpligen vägas upp med viktvariabeln. Objekt med partiellt bortfall bör också exkluderas såvida man inte vill behandla bortfallskoden som ett speciellt variabelvärde. I AID-analys kan man nämligen låta bortfallskoden hos en fri oberoende vaiabel vara en egen kategori. Detta innebär att vi får fler poster i datamaterialet, men att vissa tolkningsproblem uppstår. Detta exemplifieras i analys av HINK 1972, se avsnitt 6.1.2 där det t ex visar sig att gifta, skilda och änklingar har en medelinkomst på 33 300 kr jämfört med ogifta och ej svarande med 28 700 kr.

5 Några specialiteter

Som jag hoppas har framgått av beskrivningen är metodens egenskaper nära knutna till möjligheterna i dataprogrammet. I versionen AID 3 finns utöver den grundversion av metoden som beskrivits hittills en rad specialiteter. Nedan beskrivs dessa ganska kortfattat. Jag har valt att ibland benämna dem med det engelska namnet för att underlätta kopplingen med dataprogrammet. Vi tar i detta sammanhang också upp THAID-analys. Det speciella med den metoden är att beroende variabeln är en nominal- eller ordinalskalevariabel.

5.1 Look-ahead

En nackdel med grundvarianten av AID-metoden är att delningarna görs stegvis där bästa delning bara söks i varje steg för sig. Detta ger bästa delning av hela materialet i två grupper, men när dessa delats var sin gång och man har fyra grupper är det inte säkert att dessa fyra grupper är resultat av den bästa fyrdelningen. Genom att använda sig av look-ahead (se framåt!) ett steg gör programmet i första steget en delning som med hänsyn till nästföljande två delningar ger bästa fyrdelning av materialet. Look-ahead kan användas även två och tre steg framåt. Beräkningarna och minneskravet i datorn blir avsevärt mycket större med look-ahead och därmed kostnaden.

Att upptäcka interaktioner är vad metoden lovar men det gör den egentligen inte genom att studera bara ett steg i taget. Interaktion innebär att två oberoende variabler samverkar i deras "effekt på" beroende variabeln och endast med look-ahead söker vi verkligen efter interaktioner. I appendix 1 illustreras sökandet efter interaktioner med ett enkelt exempel.

5.2 Predefined splits

I dataanalys skall man inte låsa sig vid resultatet av en viss datakörning utan gärna "vrida och vända" på materialet för att få uppslag till idéer. Att starta en AID-analys med en i förväg given uppdelning i det första steget eller i några av de första stegen kan ibland vara motiverat. Detta åstadkommes genom att definiera "predefined splits", vilket ju är ungefär detsamma som att göra analyser på två eller flera delmaterial - men bekvämare. Det finns bl a två anledningar att göra om en AID-analys där givna delningar definieras:

1° om två oberoende variabler är ungefär lika bra delare i början av trädet så är det intressant att se hur båda träden ser ut.

2^{o} om två grupperingar av kategorier hos en variabel ger ungefär lika bra delningar i början av trädet så är det intressant att se hur båda träden ser ut, speciellt om den gruppering som ej delade har en naturligare tolkning.

5.3 Ranking

Att ge de oberoende variablerna olika "ranger" (prioriteter vore bättre på svenska) innebär att den eller de variabler med högsta rangen kommer i första hand i valet som delare. Endast om den eller dessa inte uppfyller kraven för delning får variabler med näst högsta rang chansen. De variabler med hög rang som ej kan dela en viss grupp får ej heller chansen att dela längre ut på den grenen i trädet. Rangen noll ges till variabler som man vill ha medelvärden, kvadratsummor m m för, men som inte skall ingå i trädet.

Ranking kan användas om man vill utnyttja viss kunskap eller vissa antaganden om kausala relationer mellan variablerna i uppbyggnaden av trädet.

5.4 Symmetri

Att begära symmetri är ytterligare ett sätt att lägga restriktioner på delningsprocessen. Det innebär att båda av ett par av grupper från samma föräldragrupp skall delas med samma oberoende variabel och tillgår så att när den ena gruppen på vanligt sätt delats så delas automatiskt även den andra på samma vis. Man kan åsidosätta det absoluta kravet på symmetri genom att specificera att delningen av andra gruppen måste vara "bättre" än en viss proportion av hur bra den bästa delningen av denna grupp är mätt med kriteriefunktionen.

5.5 "Regressionsanalys"

I alla hittills betraktade varianter av AID-metoden har vi sökt att inom grupperna få så små avvikelser kring medelvärdet som möjligt. Programskaparna har nu inte nöjt sig med bara detta mått på homogenitet.

Minstakvadratanpassa en rät linje som beskriver hur beroende variabeln samvarierar med en annan variabel (ingen av de oberoende variablerna i den egentliga AID-analysen). Skillnaderna mellan den beroende variabelns värden och linjen kallas residualer och vi bildar summan av de kvadrerade residualerna. Om vi klyver datamaterialet i två grupper kan vi i varje grupp anpassa en linje och beräkna residualkvadratsummorna. Den dataklyvning som ger minsta summan av residualkvadratsummorna accepterar vi, varefter denna variant av AID-analys går vidare, steg för steg.

5.6 Lutningsanalys

Den vanliga AID-analysen minimerar variationen inom grupper vilket är detsamma som att maximera avståndet mellan gruppernas medelvärden. Detta är principen också i "regressionsanalys" ovan.

Lutningsanalys (slopes analysis) innebär att lutningarna hos linjerna, bildade på samma sätt som i 5.5, i de två delgrupperna i en delning skall vara så olika som möjligt. Detta innebär inte att grupperna är speciellt homogena i olika avseenden, eftersom man helt bortser från medelvärdet av den beroende variabeln.

5.7 THAID-analys

När den beroende variabeln mäter egenskaper, åsikter eller attityder med hjälp av ett begränsat antal alternativ, som inte alltid har en naturlig ordning, fungerar inte AID. THAID har utvecklats för detta ändamål. Syftet är annars detsamma, nämligen att ta fram homogena grupper.

Säg att det finns m svarsalternativ på en fråga av vilka varje objekt endast svarat på ett. I hela datamaterialet har andelen p_i angett svaret i. Talen $(p_1, p_2, \ldots, p_m)$ beskriver hur materialet ser ut med avseende på denna variabel. När materialet klyvs så har ena delen profilen $(p_1^1, p_2^1, \ldots, p_m^1)$ och andra delen $(p_1^2, p_2^2, \ldots, p_m^2)$.

Det finns ett flertal möjliga sätt att ange ett mått på skillnaden mellan dessa profiler. I THAID-programmet finns två att välja bland, se Morgan och Messenger (1973).

Ett exempel på metoden ges i avsnitt 6.1.4 där de fem stora riksdagspartiernas andelar av väljarkåren motsvaras av $p_1, \ldots, p_5$.

6 När och hur kan man använda AID-analys

Det går inte att exakt säga när och hur AID-metoden bör användas i den statistiska produktionsprocessen vid SCB, dels beroende på analysmetodens egenskaper, dels beroende på variationen bland statistikprodukterna. I detta kapitel ger vi råd i tre avsnitt: Tips från bl a litteraturen, Exempel från SCB och Sammanfattning av AID-kritik.

6.1 Tips från bl a litteraturen

Här tar vi upp tre områden:

- AID-metodens tillämpningsområden, allmänt

- Vad man bör se upp med vid AID-analys

- Användning av statistiska analysmetoder i statistikproduktionen.

6.1.1 AID-metodens tillämpningsområden, allmänt

I Fielding (1977) pekas på fyra breda användningsområden av AID- metoden. Det första är som redskap för att systematiskt klargöra komplexa beroenden mellan variabler i en insamlad datamängd. En noggrann undersökning av resultaten ger förhoppningsvis en känsla för datastrukturen. I denna, liksom i de flesta andra situationer, bör man alltid ta del av övrig information som programmet ger än bara AID-trädet, t ex vilka variabler som är nära att dela och alternativa kategorigrupperingar.

AID-metodens skapare ansåg att metodens huvudsakliga användning är som redskap i en förundersökning för att formulera den "riktiga" statistiska modellen. Steg nummer två i den strategin är att samla in nya data och testa modellen.

Ett speciellt användningsområde för AID-metoden är att finna homogena grupper av objekt som har extrema värden på den beroende variabeln. I denna tillämpning intresserar man sig alltså bara för definitionen av en enda slutgrupp.

Det fjärde området är också speciellt. Det gäller att minska det stora antalet olika värden hos en beroende variabel genom att ersätta dem med slutgruppernas medelvärden.

Metoden kan förmodligen användas till även andra speciella ändamål.

6.1.2 Vad man bör se upp med vid AID-analys

Programskaparna har genom tester på olika datamaterial skaffat sig erfarenheter som kan vara värda att notera:

- När den beroende variabeln är kraftigt snedfördelad inom grupperna tenderar AID-algoritmen att dela många små grupper som tillhör fördelningens svans. Eftersom de har en stor kvadratsumma ger de också ett stort bidrag till "förklaringsgraden". Resultaten blir oftast svårtolkade. En lämplig transformation av den beroende variabeln, t ex genom logaritmering, kan i sådana situationer vara till hjälp i analysen. Om den beroende variabeln är dikotom, vilket är tillåtet, så bör den inte anta något av sina två värden i mindre omfattning än 20 %.

- Oberoende variabler med ungefär lika många objekt i varje kategori har lättare att inkluderas i AID-trädet än snedfördelade variabler. Monotona oberoende variabler som är symmetriskt fördelade i sina värden prioriteras av samma anledning. En följd av detta är att många delningar ger grupper av ungefär samma storlek snarare än grupper med stor skillnad i medelvärde.

- Variabler med många kategorier har större möjlighet att inkluderas beroende på att det finns större chans att finna en bra kombination av gruppstorlek och medelvärdesdifferens jämfört med fallet med få kategorier. Med signifikansresonemanget från avsnitt 3.2 skulle man ta hänsyn till denna "stora risk för slumpmässiga delningar". Man bör eftersträva att de oberoende variablerna får kategoriindelningar som ger alla oberoende variabler samma chans a priori som delare och helst hålla antalet kategorier lågt. Utnyttja, om det passar i analysen, möjligheten att definiera variabler som monotona.

- Slutligen en generell tumregel: det behövs minst tusen observationer för en AID-analys.

6.1.3 Användningen av statistiska analysmetoder i statistikproduktionen

Utnyttjandet av statistiska analysmetoder vid SCB har ökat på senare år men har inte den omfattning, som det skulle kunna ha. Holgersson (1979) urskiljer tre användningsområden:

I Användning av statistisk analysmetodik inom statistikproduktionen

II Användning i statistikproduktionen i syfte att ge konsumenterna mer information

III Användning för att ge konsumenter och uppdragsgivare bättre beslutsunderlag.

AID-metoden passar mycket bra för det första syftet t ex vid stratifiering och vid val av korstabeller. I dessa fall publicerar man AID-trädet endast i en teknisk rapport och i huvudpublikationen informerar man om att metoden använts.

Om man vet att konsumenterna rätt kan tolka en AID-analys eller om man har möjlighet att ge tillräckliga tolkningsråd kan en AID-analys publiceras i speciella PM eller bilagor. Vi bör dock inte ha AID-träd i den löpande statistikredovisningen från SCB eftersom metoden inte ger resultat som lämpar

sig för jämförelser över tiden.

AID-metoden kan för enskilda statistikkonsumenter ge ett efterfrågat beslutsunderlag. Det åligger då oss som producenter att deklarera varans kvalitet, speciellt att klargöra de begränsade tolkningsmöjligheterna.

6.2 Exempel från SCB

Det finns några intressanta tillämpningar av AID vid SCB. Vi börjar med HBU där metoden använts för efterstratifiering. I HINK 72 användes AID för att dela upp hela populationen av anställda personer i redovisningsgrupper. I en konsultation åt Domstolsverket kombinerades regressionsanalys och AID i en dataanalys. Slutligen visar vi en tillämpning av THAID-analys i PSU.

6.2.1 HBU

En ovanlig men passande tillämpning av AID görs i HBU (Hushållsbudgetundersökningen). Man har samlat in ett stort urval, och naturligtvis fått ett bortfall om vilket man dock känner vissa bakgrundsfaktorer som hushållsstorlek, hushållsföreståndarens ålder och region. Efterstratifiering minskar inverkan av bortfallsfelet. Indelningen i efterstrata skall göras så att svarsfrekvensen varierar mellan efterstrata samtidigt som efterstrata är homogena. Detta problem passar för AID-metoden. Resultatet hade dock en låg förklaringsgrad men ansågs ändå ge tips om efterstratifieringen;
det finns under givna förutsättningar knappast en bättre indelning än den som programmet gav. Se Lindström (1979).

6.2.2 HINK 72

I en bilaga i redovisningen av inkomstfördelningsstudien 1972 (SoS 1976) redovisas en studie av anställdas löner med multivariata analysmetoder. Här används AID inledningsvis för att ge en uppfattning om vilka bakgrundsfaktorer som har stor betydelse för lönedifferenser bland heltids- resp deltidsanställda. Studien fortsätter därefter med regressions-

analys av i första hand sambandet ålder-inkomst i var och en av åtta grupper som AID-analysen gett. Dessa grupper är män resp kvinnor bland akademiker/företagsledare, lägre tjänstemän,kameralt/kommersiellt anställda samt arbetare.

6.2.3 Handläggningstider

I ett uppdrag åt Domstolsverket skulle handläggningstiderna vid tingsrätterna analyseras (Norberg 1978). Tiden i rättssalen, förhandlingstiden, är en del av denna och är därför en bra förklarande variabel. Andra förklarande variabler är t ex förekomst av tolk, antal åtalade, erkända och ej erkända brott, antal försvarare, målsägande och hörda m m. Den linjära regressionsmodellen är inte helt lämplig i utforskandet av datamaterialet, eftersom den förutsätter, att skillnaden i handläggningstid mellan t ex ett och två ej erkända brott är lika stor som mellan tio och elva. Vi skulle vilja använda AID-metodens flexibla egenskaper, men det kräver att variabeln förhandlingstid måste klassindelas och därmed förloras information och förklaringsgrad. Metoderna kombinerades i följande dataanalys för att plocka fram de goda egenskaperna hos respektive metod.

1° en rät linje av handläggningstiden på förhandlingstiden minstakvadratanpassades och sedan skrevs residualerna ut på datafilen tillsammans med de andra variabelvärdena.

2° en AID-analys gjordes med dessa residualer som beroende variabel. Den första delningen bestämdes i förväg i korta resp långa mål med avseende på variabeln förhandlingstid. Eftersom den beroende variabeln och förhandlingstiden var okorrelerad gav denna delning inget förklaringsbidrag.

3° AID-analysen fortsattes med de övriga oberoende variablerna.

I denna tillämpning hade de övriga oberoende variablerna ett mycket litet förklaringsvärde i analysen.

6.2.4 PSU

THAID-metoden har prövats på PSU (Partisympatiundersökningen) och rapporterats i Metodinformation nr 76:1 i serien Statistiska metoder. Både parti- och partiblockssympatier har analyserats och de två tillgängliga kriteriefunktionerna har prövats. Som förklarande variabler användes kön, ålder m m från RTB (Registret över totalbefolkningen) samt en socioekonomisk variabel. Ur ett av träddiagrammen kunde bl a följande utläsas om de svarande i urvalet:

- Centerpartiet har sin största andel 81 % bland manliga jordbrukare med sammanlagd inkomst under 40 000 kr

- Folkpartiets största andel 23 % finns bland tjänstemän, företagare och jordbrukare, äldre än 50 år, med en sammanlagd inkomst över 40 000 kr dock icke boende i fyrstadskretsen, västsverige (inklusive Göteborg) eller Stockholms kommun.

- Bland personer med ovannämnda egenskaper dock boende i de tre uppräknade områdena har moderata samlingspartiet sin största andel 49 %.

- Arbetare över 60 år med en sammanlagd inkomst mellan 10 000 och 20 000 kr eller över 30 000 kr är socialdemokraternas starkaste grupp 81 %.

- VpK har sin största andel 22 % bland facklärda och studerande, under 40 år och med en sammanlagd inkomst under 20 000 kr.

6.3 Sammanfattning av AID-kritik

Vid SCB har vi inte gjort många fler AID-analyser än vad som exemplifierats i denna handledning. Vi har heller inte nåtts av någon speciell kritik för vårt relativt försiktiga utnyttjande. Utanför SCB och särskilt i USA finns många publicerade

tillämpningar av metoden. Många har dragit nytta av metodens flexibilitet i sökandet efter hypoteser men det finns många som fått mycket kraftig kritik för missbruk. Vi tar inte upp enskilda exempel, vilka emellertid är lätta att finna, utan försöker kort sammanfatta kritiken:

- AID-träd skall inte publiceras utan en fullständig verbal tolkning, annars kan de tas som ett bevis på vissa slutsatser som den okritiske läsaren drar.

- Använd minst 1000 observationer. Det finns en tillämpning med 70 observationer och 10 oberoende variabler som fått mördande kritik. 70 observationer kan vara ett minimum för en vanlig tvåvägs korstabell. Undantag för regeln är om man använder AID-programmet för att t ex isolera en extremgrupp av objekt.

- Korrelation mellan oberoende variabler medför att när en av dem inkluderas i trädet så används de andra kanske inte alls trots att de är korrelerade med den beroende variabeln från början. Metoden "gömmer" i st f "upptäcker" ett beroende. Ser man inte upp med detta finns risk för att man formulerar konstiga hypoteser.

7 Litteratur

Litteraturlistan är delad i 6 bitar avseende olika områden.

Beskrivning av AID-metoden

Sonquist J A, Multivariate Model Building. The Validation of a Search Strategy. Institute for Social Research. The University of Michigan, Ann Arbor, Michigan. 1971.

Sonquist J A, Baker E L, Morgan J N, Searching for Structure. Institute for Social Research, The University of Michigan, Ann Arbor, Michigan, 1973

Fielding A, Binary Segmentation: The Automatic Interaction Detector and Related Techniques for Exploring Data Structure, Ingår i O'Muircheartaigh C A, Payne C, Exploring Data Structures, John Wiley & Sons, 1977.

Beskrivning av AID-programmet

Sonquist (1973), se ovan

Särtryck ur manualen för Osiris III version 2.

Beskrivning av THAID-metod och program

Morgan J N, Messenger R C, THAID A Sequential Analysis Program for the Analysis of Nominal Scale Dependent Variables. Institute for Social Research, The University of Michigan, Ann Arbor, Michigan, 1973

Särtryck ur manualen för Osiris III version 2

Allmänt om statistisk analys

Holgersson M, Resursfördelning i statistikproduktionen, Kursmaterial till SCB-kurs, 1979-11-08

Wahlström S, Lindström H, Några synpunkter på användning av regressionsteknik vid samhällsstatistiska undersökningar. II, Statistisk tidskrift 1972:5

Wretman J H, Analys av Surveydata, 1980

"Signifikansproblemet" i AID-analys

Avén A, Förslag till lösning av signifikansproblemet i AID-analys, Forskningsrapport, Statistiska Institutionen Stockholms Universitet, Stockholm, 1974

Kass G V, Significance Testing in Automatic Interaction Detection (A.I.D), Applied Statistics 24, No. 2 1975

Norberg A, Bestämning av kritiskt värde för kriteriefunktionen i AID-analys, 1980

Scott A J, Knott M, An Approximate Test for Use with AID, Applied Statistics 25 nr 2 1976

Tillämpningar

Barns och personals närvaro och frånvaro i daghem 1977, Statistiska meddelanden S 1979:1, SCB, 1979 innehållande i bilaga; Regressionsanalys av sjukfrånvaro hos barn på daghem

Elofsson G, Larsson U-B, Anställdas löner studerade med multivariata analysmetoder, Bilaga 1 till Inkomstfördelningsstudien 1972, SoS, SCB, 1976

Hedström L, Sollander S, THAID-analys av ett PSU-material, Metodinformation Nr 76:1 i serien Statistiska metoder, SCB 1976

Lindström, H, Hushållsbudgetundersökningen 1978. Metodplanering för estimation, bortfallskompensation och variansberäkning, Metodproblem i individ- och hushållsstatistik Nr 2, SCB, 1979

Norberg A, Residualer i regressionsanalys och AID-analys, Domstolsverkets handläggningstider, rapport nr 18, SCB, 1978-08-02

Att automatiskt upptäcka interaktioner

1 Syfte

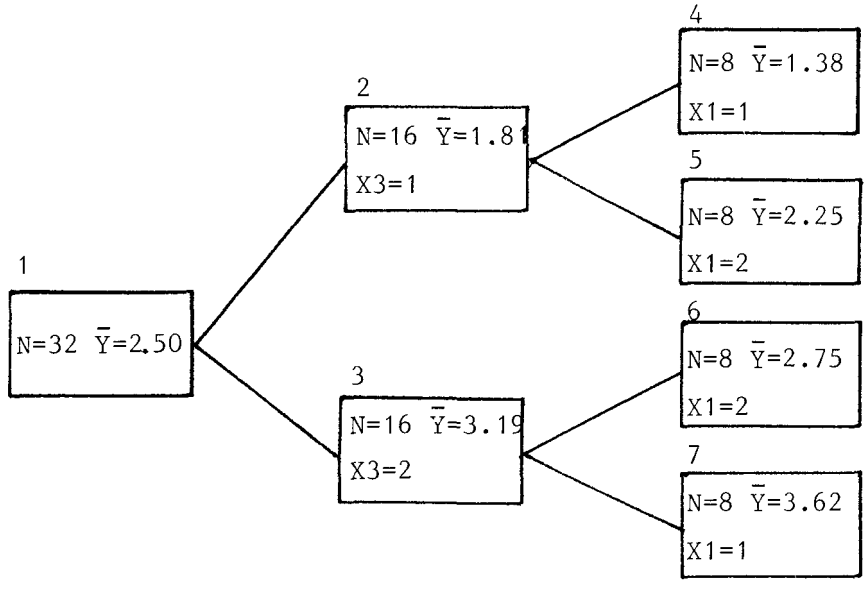
I denna korta PM ges ett exempel på AID-metodens begränsning när det gäller att upptäcka interaktioner, vilket borde vara metodens huvudsyfte (AID = Automatic Interaction Detector). För att verkligen upptäcka interaktioner bör man utnyttja look-ahead-möjligheten vilken emellertid är dyrbar när man har många variabler, klassindelningar och observationer.

2 Data

Data för analysen (redovisas på sid 3) omfattar 32 poster med variablerna Y (den beroende variabeln) och X1, X2 samt X3 (prediktorer). X3 har största korrelationen med Y men X1 och X2 har tillsammans ett exakt förhållande till Y dvs det finns interaktion mellan X1 och X2 i sambandet med Y.

3 AID-analys utan look-ahead

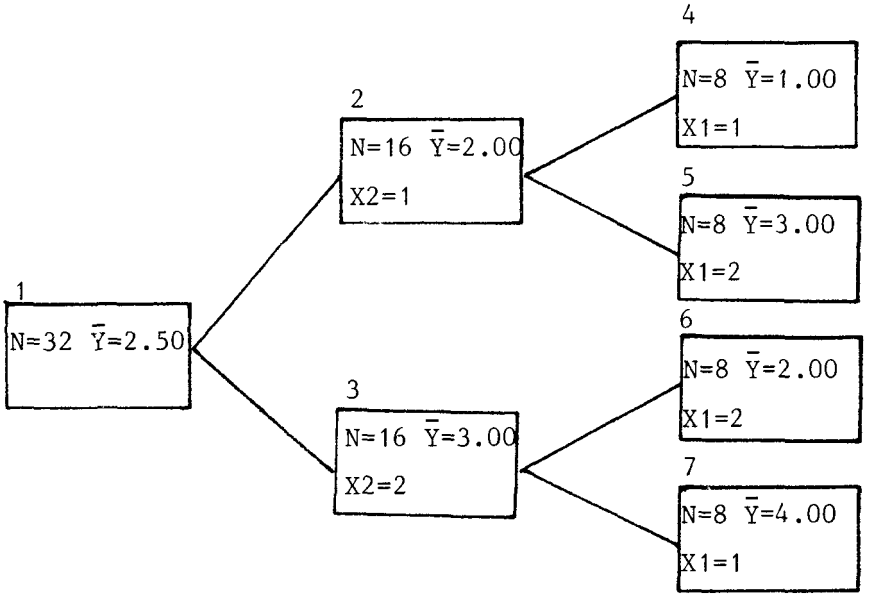
Första delningen görs efter X3 som ger större medelvärdes-differenser än vad X1 eller X2 ger. Analysen bryts efter tre steg med fyra slutgrupper.



Detta träd har 53 % förklaringsgrad. Om analysen fortsatt med uppdelning av de fyra slutgrupperna efter X2 så skulle förklaringsgraden blivit 100 %. Interaktionen mellan X1 och X2 upptäcks i detta fall med viss möda, men med riktiga data kan den vara omöjlig att se.

4 AID-analys med look-ahead

Med look-ahead ser programmet framåt och analyserar delningar av de grupper som skall bildas i analyssteget. Med look-ahead ett steg framåt på våra data upptäcks interaktionen och slutgrupperna saknar varians i Y.



Detta träd har naturligtvis förklaringsgraden 100 %.

5 Slutsats

AID är en av de grundligaste metoderna vi har för att göra explorativ dataanalys. Exemplet visar emellertid att det skulle krävas den dyrbara varianten med look-ahead för att verkligen upptäcka interaktioner.

Data för experiment med AID-analys

Y	X1	X2	X3
1	1	1	1
1	1	1	1
1	1	1	1
1	1	1	1
1	1	1	1
1	1	1	1
1	1	1	1
1	1	1	2
2	2	2	1
2	2	2	1
2	2	2	1
2	2	2	1
2	2	2	1
2	2	2	1
2	2	2	2
2	2	2	2
3	2	1	1
3	2	1	1
3	2	1	2
3	2	1	2
3	2	1	2
3	2	1	2
3	2	1	2
3	2	1	2
4	1	2	1
4	1	2	2
4	1	2	2
4	1	2	2
4	1	2	2
4	1	2	2
4	1	2	2
4	1	2	2

Program och resultatlistor

Programmet på nästa sida avser exemplet i avsnitt 2.4. I jobsteg nummer 1 skapas ett temporärt datasätt med standardprogrammet DBCRED som är effektivt vid selektering av poster. Detta steg är ej nödvändigt eftersom man även i Osirisprogrammet kan selektera poster.Observera att i Osirisprogrammet anges datasättets namn i EXEC-kortet.

Variablernas namn och lägen på datafilen deklareras efter kortet ÅDICT sist i programmet. I exemplet läses fler variabler in än vad som används i analysen.

I ett block efter kortet ÅRECODE görs vissa transformationer och aritmetiska operationer på variablerna. Där skapas också några helt nya variabler.

ÅSETUP är inledningen till styrkorten för den analys som skall göras. Därefter följer

+ selekteringskort som gäller för hela körningen
+ en titel
+ ett kort där vissa för oss ointressanta parametrar kunnat anges
+ variabellistan, dvs de variabler som verkligen skall användas
+ en bunt kort som preciserar en analys. Flera sådana buntar får finnas. Varje bunt består av
- selektering av poster
- en titel
- precisering av den beroende variabeln och eventuell viktvariabel
- två kort för precisering av de oberoende variablerna
- precisering av stoppkriterierna.

Körningen av detta program tog 1 minut och 18 sekunder CPU-tid (och kostade 480 kronor). På sidorna 3-7 presenteras några viktiga resultatsidor med tolkningsråd.

```
//ZSSTMAN JOB (446100044404),'NORBERG P/STM',
// NOTIFY=ZSSTMAN,MSGCLASS=R
/*ROUTE  PRINT RMT1
//STEG1     EXEC DBCRED
//REGIN DD DSN=C573.MV2S.BARNMP.T77.BIA0913,DISP=SHR
//REGUT DD DSN=&&TEMP,DISP=(,PASS),SPACE=(TRK,(75,5),RLSE),
//       UNIT=DISK30,DCB=(RECFM=FB,LRECL=170,BLKSIZE=3400)
//SYSPRINT DD SYSOUT=A
//SYSIN     DD *
MTDUPL
SA (14,1)¬ *
SA (18,1)¬ *
SA (18,1)¬4*
SA (19,3)¬999*
SA (22,1)¬ *
SA (55,2)>09*
SA (55,2)<76*
SA (151,3)<200
/*
//STEG2     EXEC OSIRIS,TIME=2,DA='&&TEMP'
//FT06F001 DD  SYSOUT=A,COPIES=1
//SETUP DD *
ÅRUN AID3
ÅRECODE
       R1=77-V1
       IF V10 EQ 0 THEN R2=99 ELSE R2=V10
       R3=100*V11/R2
       R4=V13/20
       R4=TRUNC(R4)
       R5=BRAC(V4,0=0,1=1,2-5=2,6-10=3,11-20=4,21-998=5)
ÅSETUP
EXCLUDE V10=00-09 OR V10=76-99  *
AIDANALYS AV SJUKFRÅNVARON VID DAGHEMMEN
*
V3,V5,V8,V12,R1,R3,R4*
INCLUDE R1=01-09 AND R3=000-100 *
BARN UNDER TIO ÅR
YVAR=R3 WEIGHT=V12 *
PRED=(R1),M,MAXCLASS=9 *
PRED=(V3,V5,V8,R4),F,MAXCLASS=8,END *
MAX=15,REDUCIBILITY=(.3,0,0) *
ÅDICT
   3   1  13   1   1
T   1 FÖDÅR                        13   20
T   2 KÖN                          17   10
T   3 BOSTAD                       18   10
T   4 AVSTÅND                      19   30
T   5 FÄRDSÄTT                     22   10
T   6 AVGFRÅN                      41   10
T   7 SYSKON                       49   10
T   8 REGION                       52   10
T   9 AVDTYP                       53   10
T  10 ÖVNÄRV                       55   20
T  11 SJUKF                        59   20
T  12 VIKT                        114   60
T  13 PERSTÄTH                    151   30
/*
```

BARN UNDER TIO ÅR

5957 OBSERVATIONS READ AFTER GLOBAL FILTER

Y AVERAGE är det ovägda medelvärdet för de 5957 observationer som passerat selekteringsvillkoren för hela körningen.

```
Y AVERAGE =            1.044267E 01
STANDARD DEVIATION =   1.254667E 01
BOUNDARIES =  -5.229068E 01   7.317601E 01
```

I tabellen listas 21 observationer som har extrema Y-värden (OUTLIERS) vilket innebär att de ligger mer än fem standardavvikelser från medelvärdet. Om inget annat anges blir de kvar i datamängden.

```
   1. ID =        79,  Y =  7.900000E 01
   2. ID =       100,  Y =  1.000000E 02
   3. ID =        80,  Y =  8.000000E 01
   4. ID =        90,  Y =  9.000000E 01
   5. ID =        79,  Y =  7.900000E 01
   6. ID =       100,  Y =  1.000000E 02
   7. ID =        80,  Y =  8.000000E 01
   8. ID =        79,  Y =  7.900000E 01
   9. ID =        76,  Y =  7.600000E 01
  10. ID =        84,  Y =  8.400000E 01
  11. ID =        87,  Y =  8.700000E 01
  12. ID =        80,  Y =  8.000000E 01
  13. ID =        92,  Y =  9.200000E 01
  14. ID =        92,  Y =  9.200000E 01
  15. ID =        92,  Y =  9.200000E 01
  16. ID =        79,  Y =  7.900000E 01
  17. ID =       100,  Y =  1.000000E 02
  18. ID =        92,  Y =  9.200000E 01
  19. ID =        95,  Y =  9.500000E 01
  20. ID =       100,  Y =  1.000000E 02
  21. ID =        77,  Y =  7.700000E 01
```

4 observationer uppfyller ej villkoren i andra selekteringskortet.

```
5953 CASES INCLUDED IN THE ANALYSIS
        4 FILTERED (LOCAL/SUBSET SELECTOR)
        0 MISSING DATA CASES
       21 OUTLIERS INCLUDED
        0 INVALID PREDICTOR VALUES
```

Det vägda medelvärdet på variabeln Y är 10.43447, viktsumman är 73 381.

```
5953 SAMPLE OBSERVATIONS - WITH TOTALS
     WEIGHTS =                    7.338100E 04
     DEPENDENT VARIABLE (Y) =     7.656920D 05    AVERAGE =    1.043447E 01
     Y-SQUARED =                  1.945003D 07    VARIANCE =   1.562035E 02
```

```
                    STAGE  1 OF THE ANALYSIS
                  BEST SPLIT BASED ON MEANS
             0-STEP LOOKAHEAD WITH 1 FORCED SPLITS

SPLITTING CRITERIA -
    MAXIMUM NUMBER OF SPLITS =  15
    MINIMUM % OBSERVATIONS IN A GROUP = 25
    %AGE OF TOTAL SS N SPLITS MUST EXPLAIN =  0.3(N=1),
    PRINT   CASES OUTSIDE  5.0 STANDARD DEVIATIONS OF PARENT GROUP MEAN
```

Vi har fem prediktorer, oberoende variabler. De som omkodats eller skapats under ÅRECODE får inget namn utan har bara sitt variabelnummer, variabel -1 är Ålder och variabel -4 är Personaltäthet.

```
       5 NON-RANKED PREDICTORS SPECIFIED
          PREDICTOR          VARIABLE NUMBER   TYPE   MAX CLASS    RANK
     1    RECODED VARIABLE              V  -1    M         9          1
     2    BOSTAD                        V   3    F         8          1
     3    FÄRDSÄTT                      V   5    F         8          1
     4    REGION                        V   8    F         8          1
     5    RECODED VARIABLE              V  -4    F         8          1

    WEIGHTED Y VARIABLE   -3  RECODED VARIABLE          SCALED BY  1.0E 00

            1 CANDIDATES -    GROUP       SS
                                 1   1.146044E 07
```

Totala kvadratsumman, TSS, är 11 460 440.

```
              ATTEMPT SPLIT ON GROUP   1 WITH SS = 1.146044E 07
```

LOOKAHEAD TENTATIVE PARTITION

Här börjar första steget. Hela materialet kallas grupp 1 och har dessa summor m m.

SPLIT ATTEMPT ON GROUP 1 WITH N = 5953, SS = 1.146044E 07
SUMS- W = 7.338100E 04, Y = 7.656920E 05, YSQ = 1.945002E 07, X =

PREDICTOR RECODED VARIABLE
9 NON-EMPTY CLASSES 1 2 3 4 5 6 7 8 9

PARTITION		N	WEIGHT	Y-MEAN	Y-VARIANCE		X-MEAN	X-VARIANCE		SLOPE	BSS
BETWEEN	1	281	1.90500E 03	2.28399E 01	3.48147E 02	0.0	0.0	0.0	0.0		
AND	2	5672	7.14760E 04	1.01038E 01	1.46909E 02	0.0	0.0	0.0	0.0		3.00982E 05
BETWEEN	2	1255	8.43900E 03	1.84734E 01	2.57042E 02	0.0	0.0	0.0	0.0		
AND	3	4698	6.49420E 04	9.38984E 00	1.33636E 02	0.0	0.0	0.0	0.0		6.16231E 05
BETWEEN	3	2524	1.88140E 04	1.59052E 01	2.24894E 02	0.0	0.0	0.0	0.0		
AND	4	3429	5.45670E 04	8.54824E 00	1.18673E 02	0.0	0.0	0.0	0.0		7.57217E 05
BETWEEN	4	3465	3.15640E 04	1.36633E 01	2.02123E 02	0.0	0.0	0.0	0.0		
AND	5	2488	4.18170E 04	7.99730E 00	1.07775E 02	0.0	0.0	0.0	0.0		5.77454E 05
BETWEEN	5	4338	4.59190E 04	1.23493E 01	1.86641E 02	0.0	0.0	0.0	0.0		
AND	6	1615	2.74620E 04	7.23272E 00	8.89830E 01	0.0	0.0	0.0	0.0		4.49882E 05
BETWEEN	6	5272	6.18390E 04	1.10468E 01	1.66491E 02	0.0	0.0	0.0	0.0		
AND	7	681	1.15420E 04	7.15361E 00	8.84421E 01	0.0	0.0	0.0	0.0		1.47427E 05

BEST SPLIT ON PREDICTOR -1 = 7.572170E 05 AFTER CLASS 3

Största BSS för Ålder (det står prediktor -1 på sista raden i detta block) är $7.6 \cdot 10^5$ vilket ger $\lambda = 7.6 \cdot 10^5 / 1.1 \cdot 10^7 = 0.066$. Detta värde erhålls när vi delar mellan 3 och 4 år.

PREDICTOR BOSTAD
3 NON-EMPTY CLASSES 1 3 2

PARTITION		N	WEIGHT	Y-MEAN	Y-VARIANCE		X-MEAN	X-VARIANCE		SLOPE	BSS
BETWEEN	1	2397	3.04330E 04	8.69645E 00	1.13891E 02	0.0	0.0	0.0	0.0		
AND	3	3556	4.29480E 04	1.16660E 01	1.82569E 02	0.0	0.0	0.0	0.0		1.57071E 05
BETWEEN	3	3716	4.60820E 04	9.40280E 00	1.33142E 02	0.0	0.0	0.0	0.0		
AND	2	2237	2.72990E 04	1.21760E 01	1.90377E 02	0.0	0.0	0.0	0.0		1.31840E 05

BEST SPLIT ON PREDICTOR 3 = 1.570710E 05 AFTER CLASS 1

Bostad är en fri prediktor. Prediktorns värden sorteras i ordning efter medelvärde på Y, dvs 1,3,2. Bästa delning är mellan 1 och 3.

PREDICTOR FÄRDSÄTT
6 NON-EMPTY CLASSES 6 2 4 3 1 5

PARTITION		N	WEIGHT	Y-MEAN	Y-VARIANCE		X-MEAN	X-VARIANCE		SLOPE	BSS
BETWEEN	2	3197	3.86490E 04	9.47039E 00	1.34612E 02	0.0	0.0	0.0	0.0		
AND	4	2756	3.47320E 04	1.15073E 01	1.78100E 02	0.0	0.0	0.0	0.0		7.58960E 04
BETWEEN	4	3475	4.20700E 04	9.47980E 00	1.33754E 02	0.0	0.0	0.0	0.0		
AND	3	2478	3.13110E 04	1.17172E 01	1.83561E 02	0.0	0.0	0.0	0.0		8.98610E 04
BETWEEN	3	3633	4.40040E 04	9.56811E 00	1.37626E 02	0.0	0.0	0.0	0.0		
AND	1	2320	2.93770E 04	1.17322E 01	1.81292E 02	0.0	0.0	0.0	0.0		8.25020E 04

BEST SPLIT ON PREDICTOR 5 = 8.986100E 04 AFTER CLASS 4

PREDICTOR REGION
2 NON-EMPTY CLASSES 2 1

PARTITION		N	WEIGHT	Y-MEAN	Y-VARIANCE		X-MEAN	X-VARIANCE		SLOPE	BSS
BETWEEN	2	3356	3.73280E 04	8.77617E 00	1.25536E 02	0.0	0.0	0.0	0.0		
AND	1	2597	3.60530E 04	1.21514E 01	1.82215E 02	0.0	0.0	0.0	0.0		2.08930E 05

BEST SPLIT ON PREDICTOR 8 = 2.089300E 05 AFTER CLASS 2

PREDICTOR RECODED VARIABLE
7 NON-EMPTY CLASSES 4 3 7 5 2 0 1

PARTITION		N	WEIGHT	Y-MEAN	Y-VARIANCE		X-MEAN	X-VARIANCE		SLOPE	BSS
BETWEEN	4	330	4.94800E 03	7.14329E 00	9.19881E 01	0.0	0.0	0.0	0.0		
AND	3	5623	6.84330E 04	1.06724E 01	1.60027E 02	0.0	0.0	0.0	0.0		5.74710E 04
BETWEEN	3	1221	1.92000E 04	7.18146E 00	9.39104E 01	0.0	0.0	0.0	0.0		
AND	7	4732	5.41810E 04	1.15872E 01	1.73227E 02	0.0	0.0	0.0	0.0		2.75175E 05
BETWEEN	7	1252	1.97360E 04	7.19811E 00	9.35700E 01	0.0	0.0	0.0	0.0		
AND	5	4701	5.36450E 04	1.16251E 01	1.74004E 02	0.0	0.0	0.0	0.0		2.82765E 05
BETWEEN	5	1367	2.12960E 04	7.35819E 00	9.45932E 01	0.0	0.0	0.0	0.0		
AND	2	4586	5.20850E 04	1.16923E 01	1.75972E 02	0.0	0.0	0.0	0.0		2.83937E 05
BETWEEN	2	3902	5.52560E 04	9.26962E 00	1.33707E 02	0.0	0.0	0.0	0.0		
AND	0	2051	1.81250E 04	1.39857E 01	2.08138E 02	0.0	0.0	0.0	0.0		3.03548E 05
BETWEEN	0	3963	5.58150E 04	9.30976E 00	1.34727E 02	0.0	0.0	0.0	0.0		
AND	1	1990	1.75660E 04	1.40082E 01	2.07754E 02	0.0	0.0	0.0	0.0		2.94949E 05

BEST SPLIT ON PREDICTOR -4 = 3.035480E 05 AFTER CLASS 2

RECODED VARIABLE YIELDS MAXIMUM BSS = 7.572170E 05

```
0-STEP LOOKAHEAD TO SPLIT GROUP   1, TOTAL BSS =  7.572170E 05
       1. SPLIT GROUP   1 ON PREDICTOR   -1, BSS =  7.572170E 05

              ***** PARTITION OF GROUP   1 *****

     MAXIMUM ELIGIBLE BSS AT EACH STEP          MAXIMUM TOTAL BSS (LOOKAHEAD)
   1.SPLIT   1 ON V  -1 BSS= 7.57217E 05    SPLIT   1 ON V  -1 BSS = 7.57217E 05
   PE(1)= 3.438E 04, TOTAL= 7.57217E 05                     TOTAL= 7.57217E 05

SPLIT GROUP   1 ON RECODED VARIABLE          V  -1

EXTREME CASES LYING OUTSIDE THE INTERVAL (-5.205618E 01, 7.292513E 01)
        Y          W          WY          V  -3
       79        2.0    1.58000E 02         79
      100        2.0    2.00000E 02        100
       80        2.0    1.60000E 02         80
       90        2.0    1.80000E 02         90
       79        9.0    7.11000E 02         79
      100        9.0    9.00000E 02        100
       80        9.0    7.20000E 02         80
       79        9.0    7.11000E 02         79
       76        9.0    6.84000E 02         76
       84        9.0    7.56000E 02         84
       87        9.0    7.83000E 02         87
       80        9.0    7.20000E 02         80
       92        6.0    5.52000E 02         92
       92        6.0    5.52000E 02         92
       92        6.0    5.52000E 02         92
       79        6.0    4.74000E 02         79
      100        6.0    6.00000E 02        100
       92        6.0    5.52000E 02         92
       73        6.0    4.38000E 02         73
       73       22.0    1.60600E 03         73
       95       22.0    2.09000E 03         95
      100       22.0    2.20000E 03        100
       77       14.0    1.07800E 03         77
    GROUP   2 WITH  3429 OBSERVATIONS FROM  6 CLASSES =  4  5  6  7  8  9
            W= 5.45670E 04 Y= 4.66452D 05 YSQ= 1.04611D 07 X=
    GROUP   3 WITH  2524 OBSERVATIONS FROM  3 CLASSES =  1  2  3
            W= 1.88140E 04 Y= 2.99240D 05 YSQ= 8.98893D 06 X=

     2 CANDIDATES -     GROUP      SS
                          2   6.473738E 06
                          3   4.229466E 06

       ATTEMPT SPLIT ON GROUP   2 WITH SS = 6.473738E 06
```

Sammanfattningen av första steget är att prediktorn -1 hade största BSS.

En ny lista över outliers.

Definitionen av de nya grupperna 2 och 3. W är viktsumman och Y är viktade summan av Y-värden.

Valet mellan grupp 2 och 3 för nästa steg faller på 2:an eftersom den har större kvadratsumma. (Det visar sig emellertid att delningen av grupp 3 gav större variansförklaring)

```
                                GROUP SUMMARY TABLE
                          15 GROUPS OF WHICH  8 ARE FINAL

GROUP   1 , N =    5953, SUM W =  7.338100E 04
    Y MEAN= 1.043447E 01, VARIANCE= 1.562033E 02,   SS(L)/TSS= 1.000, BSS/TSS= 0.066
    SPLIT ON RECODED VARIABLE          , BSS(L) =  7.572170E 05  INTO
          2 WITH CLASSES   4  5  6  7  8  9
          3 WITH CLASSES   1  2  3

GROUP   2 , N =    3429, SUM W =  5.456700E 04
    Y MEAN= 8.548243E 00, VARIANCE= 1.186730E 02,   SS(L)/TSS= 0.565, BSS/TSS= 0.008
    SPLIT ON RECODED VARIABLE          , BSS(L) =  9.567700E 04  INTO
          4 WITH CLASSES   6  7  8  9
          5 WITH CLASSES   4  5

GROUP   3 , N =    2524, SUM W =  1.881400E 04
    Y MEAN= 1.590518E 01, VARIANCE= 2.248934E 02,   SS(L)/TSS= 0.369, BSS/TSS= 0.009
    SPLIT ON RECODED VARIABLE          , BSS(L) =  1.019330E 05  INTO
          6 WITH CLASSES   2  3
          7 WITH CLASSES   1

GROUP   5 , N =    1814, SUM W =  2.710500E 04
    Y MEAN= 9.881091E 00, VARIANCE= 1.452899E 02,   SS(L)/TSS= 0.343, BSS/TSS= 0.008
    SPLIT ON REGION                    , BSS(L) =  9.174200E 04  INTO
          8 WITH CLASSES   2
          9 WITH CLASSES   1

GROUP   6 , N =    2243, SUM W =  1.690900E 04
    Y MEAN= 1.512390E 01, VARIANCE= 2.051116E 02,   SS(L)/TSS= 0.302, BSS/TSS= 0.007
    SPLIT ON FÄRDSÄTT                  , BSS(L) =  7.739900E 04  INTO
          10 WITH CLASSES   6  4  2
          11 WITH CLASSES   1  5  3

GROUP   4*, N =    1615, SUM W =  2.746200E 04
    Y MEAN= 7.232721E 00, VARIANCE= 8.898344E 01,   SS(L)/TSS= 0.213, BSS/TSS= 0.0

GROUP   9 , N =     690, SUM W =  1.226300E 04
    Y MEAN= 1.190508E 01, VARIANCE= 1.762148E 02,   SS(L)/TSS= 0.188, BSS/TSS= 0.005
    SPLIT ON RECODED VARIABLE          , BSS(L) =  5.467100E 04  INTO
          12 WITH CLASSES   0  4  3
          13 WITH CLASSES   1  2  5

GROUP  11*, N =     970, SUM W =  7.530000E 03
    Y MEAN= 1.751167E 01, VARIANCE= 2.506120E 02,   SS(L)/TSS= 0.164, BSS/TSS= 0.0

GROUP  13*, N =     553, SUM W =  9.389000E 03
    Y MEAN= 1.307328E 01, VARIANCE= 1.983479E 02,   SS(L)/TSS= 0.162, BSS/TSS= 0.0

GROUP   8*, N =    1124, SUM W =  1.484200E 04
    Y MEAN= 8.208798E 00, VARIANCE= 1.137233E 02,   SS(L)/TSS= 0.147, BSS/TSS= 0.0

GROUP  10*, N =    1273, SUM W =  9.379000E 03
    Y MEAN= 1.320684E 01, VARIANCE= 1.604977E 02,   SS(L)/TSS= 0.131, BSS/TSS= 0.0

GROUP   7 , N =     281, SUM W =  1.905000E 03
    Y MEAN= 2.283989E 01, VARIANCE= 3.481409E 02,   SS(L)/TSS= 0.058, BSS/TSS= 0.003
    SPLIT ON REGION                    , BSS(L) =  3.438956E 04  INTO
          14 WITH CLASSES   2
          15 WITH CLASSES   1

GROUP  15*, N =     140, SUM W =  1.106000E 03
    Y MEAN= 2.645117E 01, VARIANCE= 3.848101E 02,   SS(L)/TSS= 0.037, BSS/TSS= 0.0

GROUP  12*, N =     137, SUM W =  2.874000E 03
    Y MEAN= 8.088726E 00, VARIANCE= 8.559291E 01,   SS(L)/TSS= 0.021, BSS/TSS= 0.0

GROUP  14*, N =     141, SUM W =  7.990000E 02
    Y MEAN= 1.784105E 01, VARIANCE= 2.570154E 02,   SS(L)/TSS= 0.018, BSS/TSS= 0.0
```

Efter alla analyssteg sammanfattas resultatet i denna tabell. Denna kan användas som underlag för att rita upp ett AID-träd.

100*BSS/TSS TABLE FOR 0-STEP LOOK-AHEAD
15 GROUPS, 5 PREDICTORS
< INDICATES LESS THAN 1 SPLITS WERE MADE

PREDICTOR	GROUP NUMBER (* INDICATES THE GROUP IS FINAL)														
	1	2	3	5	6	4*	9	11*	13*	8*	10*	7	15*	12*	14*
-1	6.6	0.8	0.9	0.0	0.4	0.0	0.0	0.2	0.0	0.1	0.1	*****	*****	0.0	*****
3	1.4	0.5	0.7	0.5	0.6	0.1	0.2	0.1	0.1	0.2	0.1	0.1	0.0	0.1	0.0
5	0.8	0.2	0.9	0.2	0.7	0.0	0.0	0.1	0.0	0.1	0.0	0.1	0.0	0.0	0.0
8	1.8	0.7	0.8	0.8	0.5	0.1	*****	0.2	*****	*****	0.1	0.3	*****	*****	*****
-4	2.6	0.6	0.4	0.5	0.3	0.1	0.5	0.0	0.1	0.2	0.2	0.0	0.0	0.0	0.0

***** END OF ANALYSIS *****

Denna tabell sammanfattar varje steg med varje prediktors största λ-värde. Tabellen är tillsammans med ett träddiagram en komprimerad men ganska fullständig redovisning av analysen.