Rapport R13:1978

Diskriminant- och logitanalys

— en metodjämförelse

Carl-Olof Berglund
Göran Tegner
Staffan Widlert

Byggforskningen

R13:1978

DISKRIMINANT- OCH LOGITANALYS

- en metodjämförelse

Carl-Olof Berglund
Göran Tegner
Staffan Widlert

Denna rapport hänför sig till forskningsanslag 750594-6 från Statens råd för byggnadsforskning till Allmänna Ingenjörsbyrån AB, Stockholm

Nyckelord:

trafik
resor
färdmedelsval
sannolikhet
analysmetoder
statistiska metoder
diskriminantanalys
logitanalys
metodjämförelser

UDK 519:711.7
656.001.5

R13:1978

ISBN 91-540-2811-6
Statens råd för byggnadsforskning, Stockholm

LiberTryck Stockholm 1978 850636

FÖRORD

Detta forskningsprojekt har genomförts vid Allmänna Ingenjörsbyråns trafikavdelning i Stockholm. Projektledare har varit civ ing C-O Berglund och utredningsman civ ing Staffan Widlert. Pol mag Göran Tegner har medverkat som expert under projektet. Professor Åke Claesson initierade projektet och svarade för den ursprungliga uppläggningen.

INNEHÅLLSFÖRTECKNING

1 INLEDNING

Olika metoder för att analysera individers val av färdmedel har under senare år använts i Sverige. Framför allt är det diskriminantanalys och logitanalys som kommit till användning. Diskriminantanalys har använts för färdmedelsvalsstudier i Lund och Norrköping. Logitanalys har använts i Stockholm, Malmö, Uppsala och Västerås. Föreliggande undersökning syftar till att empiriskt jämföra de två analysmetoderna. Jämförelsen görs genom att använda båda metoderna på samma observationsmaterial och därigenom studera hur väl de kan förklara det observerade färdmedelsvalet.

Den ursprungliga avsikten med forskningsprojektet var att göra en logitanalys av ett undersökningsmaterial för arbetsresor i Norrköping. I ett tidigare projekt användes detta material för att utveckla färdmedelsvalsmodeller med hjälp av diskriminantanalys (1). Datamaterialet visade sig dock ha vissa svagheter som försvårade en rättvis jämförelse, varför resultaten bara medgav begränsade slutsatser (se bilaga 1). Projektet utvidgades därför till att också omfatta en diskriminantanalys av ett datamaterial från Västerås och Uppsala 1974 som samlades in och analyserades med logitanalys av Allmänna Ingenjörsbyrån AB (2). Metodjämförelserna i rapportens huvudtext baserar sig följaktligen på detta senare datamaterial.

I rapporten jämförs de två metodernas användbarhet för att beskriva individers val av färdmedel. Vi betraktar i detta sammanhang färdmedelsvalet som ett val mellan två olika alternativ (=binärt val), t ex val mellan bil och buss eller val mellan bil och "bästa" alternativa färdsätt. Det är värt att observera att metoderna även kan användas för att beskriva andra delar av trafikantens valsituation, t ex vägval, val av målpunkt etc, och att logitanalysen kan utvidgas till att beskriva val mellan flera olika alternativ. Logitmodeller har exempelvis använts för att beskriva trafikanters samtidiga val av målpunkt och färdsätt vid inköpsresor (3).

---

(1) Lindahl L & Eklind B, 1972, Val mellan bil och kollektiva färdmedel i Norrköping (Statens institut för byggnadsforskning) Rapport R 19:1972, Stockholm.

(2) Hur parkeringsanläggningars utnyttjande beror på gångavstånd, parkeringsavgift och kollektiva resmöjligheter, 1974 (Allmänna Ingenjörsbyrån AB) Stockholm.

(3) Berglund, Tegner, Widlert, 1977, Val av resmål och färdsätt vid inköpsresor - en beteendestudie (Statens råd för byggnadsforskning) Rapport R8:1977, Stockholm.

2 BESKRIVNING AV METODERNA

Logitanalys och diskriminantanalys används för att konstruera modeller som beskriver sannolikheten att välja ett visst alternativ (i vårt fall ett visst färdsätt), och hur denna sannolikhet varierar med olika förklaringsvariabler. Allmänt kan modellerna skrivas:

$$P_m = f\ (T_j,\ S_k,\ V_i)$$

där P_m = sannolikheten att en individ väljer färdsätt m

T_j = transportsystemvariabler (t ex restid, reskostnad, reskomfort)

S_k = socio-ekonomiska variabler (t ex ålder, kön, inkomst)

V_i = valvariabler, dvs de variabler som uttrycker graden av valfrihet för att välja färdmedel

Problemet är att välja ut lämpliga förklaringsvariabler och att bestämma funktionen f.

Logitanalys

Logitmodellen har utvecklats från modern beteendeteori. Hur modellen kan härledas på detta sätt visas i bilaga 4. Sambandet mellan sannolikheten för en viss individ att välja ett visst färdsätt och en kombination av de förklarande variablerna beskrivs av en S-formad kurva (en s k sigmoid). Modellen kan skrivas på följande sätt:

$$P_b = \frac{e^{L(X)}}{1 + e^{L(X)}}$$

där P_b = sannolikheten att individen väljer bil som färdmedel

$L(X)$ = en linjär funktion av variabler som förklarar färdmedelsvalet

Sambandet illusteras i figur 1:

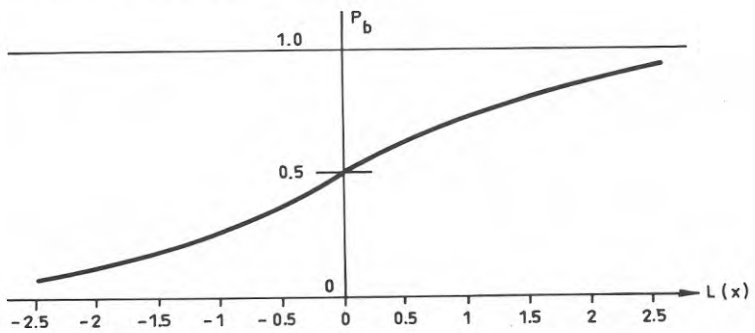

FIG 1 Grafisk beskrivning av logitmodellen

Av figuren framgår att sannolikheten att välja bil alltid kommer att ligga mellan 0 och 1. Kurvans lutning visar att känsligheten för förändringar i L(X) är störst när L(X) är noll och sannolikheten att välja bil är 0.5, dvs då konkurrensen mellan färdsätten är starkast (då färdsätten har lika stora andelar). När t ex färdsättet bil är helt överlägset betyder således exempelvis en höjning av parkeringsavgifterna mindre än när tänkbara alternativa färdsätt är helt likvärdiga i utgångsläget.

Funktionen L(X) i modellen kan skrivas:

$$L(X) = \beta_0 + \sum_{j=1}^{n} \beta_j X_j$$

där X_j = variabler som förklarar färmedelsvalet

β_0 = konstant som beror på de faktorer som ej tagits med i analysen

β_j = koefficienter för de olika variablerna

Koefficienterna i modellen bestäms med statistiska metoder så att modellen anpassar sig så väl som möjligt till ett givet datamaterial, dvs så att modellen så väl som möjligt beskriver ett verkligt observerat färdmedelsval. Mer om den binära logitmodellen kan läsas i t ex referens (2) eller (4).

Diskriminantanalys

Diskriminantanalys utvecklades ursprungligen inom biologin. Metoden används för att klassificera observationer till delpopulationer med utgångspunkt från kända variabelvärden. När vi studerar binärt färdmedelsval innebär detta att vi antar att individerna tillhör endera av två olika grupper, t ex gruppen biläkare eller gruppen bussåkare. Diskriminantanalys är således en matematisk - statistisk metod för att så säkert som möjligt klassificera individer till rätt grupp.

Diskriminantanalysens syfte är att bestämma en linjär funktion av variabler som kan användas för klassificeringen, den s k diskriminantfunktionen Z:

$$Z = \sum_{j=1}^{n} \beta_j X_j$$

---

(4) de Donnea F.X., 1971, The Determinants of transport mode choice in Dutch cities (Rotterdam University press) Rotterdam.

där X_j = variabler som "förklarar" grupptillhörigheten

β_j = koefficienter för variablerna

Vid höga värden på Z klassificeras observationen till den ena populationen, vid låga till den andra. Vanligen kan man inte finna ett Z-värde som helt skiljer grupperna åt, utan det blir en viss överlappning. Figur 2 visar hur separeringen kan se ut i praktiken.

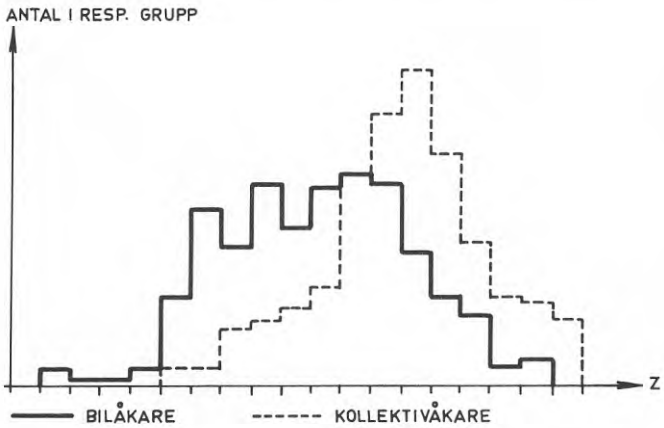

FIG 2

Som synes av figuren är överlappningen mellan grupperna i detta exempel betydande. Figuren visar att vi inte kan säga att en person med ett visst Z-värde med säkerhet tillhör den ena eller den andra gruppen. Istället kan man konstatera att det mot varje Z-värde kan anses svara en viss sannolikhet att välja respektive färdsätt. Detta stämmer också väl överens med syftet vid färdmedelsvalsstudier då målet ju normalt inte är att klassificera individer, utan snarare att förutsäga färdmedelsandelar under olika förhållanden.

Vi önskar alltså använda diskriminantfunktionen i en modell som anger sannolikheten för att välja ett visst färdsätt, dvs vi vill med en lämplig funktion översätta (transformera) varje värde på Z till en sannolikhet. Under olika antaganden kan en sådan modell få olika form. Lindahl & Eklind (1) använde en linjär sannolikhetsmodell, medan vi väljer att använda en modell med samma form som den tidigare beskrivna logitmodellen. Valet av sannolikhetsmodell behandlas inte vidare i detta sammanhang, utan den intresserade hänvisas till referens (5) och (6).

---

(5) Warner, S L, 1962, Stochastic choice of mode in urban travel: A study in binary choice (Northwestern University Press)

(6) Quarmby, D A, 1967, Choice of travel mode for the journey to work. Journal of Transport Economics an Policy, vol I, nr 3, sept. 1967.

Den valda modellen har formen

$$P_b = \frac{e^{L(Z)}}{1 + e^{L(Z)}}$$

där L(Z) = en linjär funktion av diskriminanatfunktionen, dvs en linjär funktion av förklaringsvariablerna i Z

Funktionen L väljs så att sannolikhetsmodellen ansluter sig så väl som möjligt till det observationsmaterial som använts vid beräkningen av Z. En metod för att göra detta som har angivits av Quarmby (6) beskrivs i bilaga 5. Där beskrivs även beräkningen av använda noggrannhetsmått.

Diskriminantanalysen bygger på mer restriktiva statistiska förutsättningar än logitanalysen. Förutsättningen för diskriminantanalysen är att förklaringsvariablerna i var och en av de två populationerna har en multivariat normalfördelning med olika medelvärdesvektorer men identiska varians - covariansmatriser.

3 JÄMFÖRELSE MELLAN ANALYSMETODERNA

Av den del av metodjämförelsen som redovisas i bilaga 1 framgår att samma uppsättning förklaringsvariabler ger (statistiskt sett) bäst resultat i båda modelltyperna, dvs att samma modell erhålls med båda metoderna (modellernas koefficienter har dock olika storlek).

För metodjämförelsen i detta kapitel används ett datamaterial som samlades in för projektet "Hur parkeringsanläggningars utnyttjande beror på gångavstånd, parkeringsavgift och kollektiva resmöjligheter" (2). Materialet avser arbetsresor gjorda av individer i Uppsala och Västerås med reell tillgång till bil. Den del av datamaterialet som vi använder här består av 496 observationer. Den aktuella valsituationen är valet mellan bil och bästa alternativa färdsätt. 289 personer i datamaterialet åkte bil och 207 åkte med alternativt färdsätt. Medelvärden, standardavvikelser, minimum och maximum för de aktuella variablerna visas i bilaga 2.

Som tidigare nämnts är en förutsättning för diskriminantanalysen att de två grupperna (i detta fall bilåkare respektive icke bilåkare) har lika varians-kovariansmatriser. Quarmby (6) anger som tumregel att högst ett element av tjugo i den ena matrisen får avvika med mer än faktorn två från motsvarande element i den andra matrisen. Detta krav är inte uppfyllt i vårt fall, mer än vart fjärde element avviker med mer än faktorn två. Hur mycket resultaten påverkas av denna brist på överensstämmelse med de teoretiska förutsättningarna är dock svårt att ange. Den fortsatta empiriska jämförelsen får därför bilda underlag för bedömningen av de två metoderna.

Jämförelsen görs mellan modeller med de enligt det tidigare projektet bästa förklaringsvariablerna. De erhållna modellerna visas i tabellen nedan.

Variabel X_j	Koefficient β_j (t-värde)	
	logitmodell	diskriminantmodell
Kostnadsdifferens (kostnad alt. färdsätt minus parkeringskostnad bil) kr per dag	0.1979 (3.88)	0.1709 (3.99)
Tidsdifferens exkl gångtid bil (alt. färdsätt minus bil) min. per dag	0.0190 (3.47)	0.0170 (3.55)
Gångtid bil (tid från parkering till arbete) min. per dag	-0.1044 (4.28)	-0.0990 (4.52)
Bil i tjänsten minst en gång per vecka. 1=ja, 0=nej	1.3919 (5.76)	1.3008 (6.29)
Konstant	0.1934 (0.80)	0.3024 -
χ^2	98.1504	-
pseudo R^2 respektive R^2	0.2416	0.1776

Samtliga variabler har förväntat tecken. Siffrorna inom parentes är t-värden. Detta värde är absolutbeloppet av koefficientvärdet dividerat med medelfelet i koefficientberäkningen. Värdet utgör ett test på om koefficienten är signifikant skild ifrån noll. Ju högre t-värde, desto större sannolikhet för att koefficienten verkligen är skild ifrån noll. Signifikansgränserna för testet är:

	t-värde
99.9 %	3.29
99 %	2.58
95 %	1.96

I båda modellerna är samtliga koefficienter signifikant skilda från noll på 99.9%-nivån. Diskriminantmodellens koefficienter har genomgående något högre signifikanser än logitmodellens, men skillnaden är mycket liten. Logitmodellens konstant är relativt liten och ej signifikant (för diskriminantmodellen kan konstantens t-värde ej beräknas). Eftersom konstanten fångar in det som vi inte förklarat med de övriga variablerna så önskar vi naturligtvis att den skall vara liten. Är den liten så spelar det inte heller någon roll om den är signifikant skild ifrån noll eller ej.

R^2-värdena visar att logitmodellen har något högre förklaringsgrad än diskriminantmodellen.

När tidsvärden beräknas ur de två modellerna erhålls följande resultat:

		logitmodell	diskriminantmodell
restid	kr/tim	5.76	5.97
gångtid bil	kr/tim	31.65	34.76

Skillnaderna mellan tidsvärdena från de två modellerna är mycket små, dvs metoderna ger i detta avseende likvärdiga resultat.

För att testa modellernas prognosförmåga delades observationsmaterialet i två delar och de två modellerna estimerades på den ena halvan av materialet. De på så sätt erhållna modellerna användes för att göra en prognos av hur individerna i den andra halvan av materialet skulle bete sig i nuläget. Beräkningen gjordes så att varje individs sannolikhet att åka bil beräknades separat genom att använda individens variabelvärden i modellen. Summan av samtliga individers sannolikheter är lika med prognosticerat antal biläkare:

Verkligt antal biläkare	Prognosticerat antal enligt: logitmodell	diskriminantmodell
145	142	146

Logitmodellens prognos är 2.1% för låg och diskriminantmodellens 0.7% för hög. Diskriminantmodellen ger således något bättre resultat men skillnaden är obetydlig och båda modellernas resultat är utmärkta.

Modellerna som estimerats på hela materialet användes för att beräkna antalet bilåkare i hela materialet (en test på hur väl modellerna beskriver de observationer de estimerats för). Beräkningen gjordes på samma sätt som i föregående test.

Verkligt antal bilåkare	Prognosticerat antal enligt: logitmodell	diskriminantmodell
289	289	296

Logitmodellen ger exakt rätt antal bilåkare medan diskriminantmodellen ger ett något för stort antal (2.4%). Att en modell väl återger bilandelen i observationsmaterialet är dock endast ett nödvändigt, men inte tillräckligt, krav på en prognosmodell.

> "A model can duplicate the data perfectly, but may serve no useful purpose for prediction if it represents erroneous behavioral assumptions" (Ben Akiva (9))

För att pröva modellernas klassificeringsförmåga gjordes en beräkning där individen klassificerades som bilåkare om den beräknade sannolikheten för att välja bil var större än 0.5 och som användare av alternativt färdsätt om sannolikheten var mindre än 0.5. En sådan klassificeringsberäkning kan användas för att jämföra olika modeller. Storleken på de erhållna andelarna felklassificerade kan dock inte användas som ett mått på hur bra modellerna är. Detta illustreras bäst med ett enkelt exempel. Vi tänker oss att vi har utvecklat en modell som beskriver individernas val mellan bil och buss helt perfekt. Vi tänker oss vidare att en individ i observationsmaterialet har sannolikheten 0.60 för att välja bil och 0.40 för att välja buss. Modellen beräknar då i klassificeringstesten helt riktigt sannolikheten att välja bil till 0.60 (eftersom modellen var perfekt). Trots detta är ju sannolikheten 0.40 för att individen i verkligheten åkt buss, dvs vi har 40% chans att ange individen som felklassificerad trots att modellen beskrivit valsituationen helt perfekt. Att en individ betecknas som felklassificerad innebär därför inte att modellen givit fel resultat, och absoluta storleken på andelen felklassificerade är därigenom också ointressant.

(9) Ben Akiva, M, A disaggregate direct demand model for simultaneous choice of mode and destination (International conference on Transportation Research) Belgien

I tabellen nedan visas antalet felklassificerade individer i de två grupperna (289 åkte i verkligheten bil och 207 alternativt färdmedel).

	Felklassificerade biläkare		Felklassificerade alternativåkare		Totalt	
	Antal	%	Antal	%	Antal	%
Logit-modell	62	21	88	43	150	30
Diskriminant-modell	55	19	99	48	154	31

Även klassificeringstesten visar likartade resultat för de båda metoderna. Logitmodellen felklassificerar fler biläkare medan diskriminantmodellen felklassificerar fler alternativåkare. Det totala antalet felklassificerade är något mindre för logitmodellen.

För att undersöka om modellerna skulle ge samma svar i en prognossituation beräknades slutligen med vardera modellen effekten av dels en fördubbling av de rapporterade parkeringsavgifterna i datamaterialet, dels av en fördubbling av bilrestiden. Logitmodellen angav då att bilandelen skulle sjunka med 2.7 respektive 9.3 procentenheter och diskriminantmodellen att bilandelen skulle sjunka med 2.7 respektive 9.8 procentenheter. Modellerna ger således nästan exakt samma svar på frågan om åtgärdernas effektivitet.

4 RESULTAT FRÅN ANDRA UNDERSÖKNINGAR

Tidigare svenska studier på området saknas, men jämförelser mellan diskriminantanalys och logitanalys har gjorts i USA av Watson (7) och Mc Donald (8). I båda dessa undersökningar konstaterades att samma förklaringsvariabler gav bäst resultat med båda metoderna och att modellkoefficienterna och modellerna i sig fick likvärdiga signifikanser. Mc Donald erhöll ungefär samma tidsvärden från båda modelltyperna, 1.924 cent per minut från diskriminantmodellen och 1.759 cent per minut från logitmodellen.

Testen av modellernas prognosförmåga genom att estimera modellen på hälften av datamaterialet och sedan använda den på andra hälften, gav i Watsons studie följande resultat (valsituationen som studerades var valet mellan tåg och bil):

Verkligt antal tågåkare	Prognosticerat antal enligt: logitmodell	diskriminantmodell
250	246	292

Logitmodellens prognos är 1.6% för låg och diskriminantmodellens 16.8% för hög. Diskriminantmodellen gav således här ett klart sämre resultat än logitmodellen. Även i klassificeringstesten gav diskriminantmodellen klart sämre resultat:

	Felklassificerade färdsätt 1	Felklassificerade färdsätt 2	Totalt	
	Antal	Antal	Antal	%
Logit-modell	58	146	204	33
Diskriminant-modell	218	183	401	65

Diskriminantmodellen felklassificerar dubbelt så många individer som logitmodellen. Detta är speciellt anmärkningsvärt eftersom diskriminantanalysen ursprungligen utvecklades just för att klassificera observationer. Watsons slutsats blev att diskriminantanalysen var klart underlägsen logitanalysen.

Mc Donald gjorde inte någon prognostest utan enbart klassificeringstesten (det hade knappast heller varit möjligt att dela

(7) Watson, P L, 1974, Choice of estimation procedure for models of binary choice. Regional and Urban Economics, no 4, 1964.

(8) Mc Donald, J F, Choice of estimation procedure for models of binary choice. Further Evidence.

observationsmaterialet eftersom det endast bestod av 115 observationer). Den studerade valsituationen utgjordes av bilåkande individers val mellan en avgiftsfri vanlig väg och en avgiftsbelagd men snabbare tullväg. Resultatet av klassificeringstesten blev:

	Felklassificerade som åkt tullvägen		Felklassificerade som åkt annan väg		Totalt	
	Antal	%	Antal	%	Antal	%
Logit-modell	13	28	8	12	21	18
Diskriminant-modell	14	30	8	12	22	19

Modellerna gav således likvärdiga resultat och Mc Donalds slutsats blev att diskriminantanalysen inte nödvändigtvis behöver förkastas till förmån för logitanalysen.

De två relaterade studierna har således givit motstridiga resultat.

5 SLUTSATSER

Både föreliggande studie och refererade utländska studier visar att diskriminantanalys och logitanalys ger samma resultat när det gäller urval av förklaringsvariabler. För båda de använda datamaterialen (kap. 3 och bilaga 1) gav logitanalysen något högre förklaringsgrad, men skillnaden mellan metoderna var liten. Tidsvärden erhållna ur de olika modelltyperna är mycket lika.

Diskriminantmodellen gav aningen bättre resultat när modellernas prognosförmåga testades, medan logitmodellen bättre återgav bilandelen i hela materialet och dessutom klassificerade individerna något bättre. Alla erhållna skillander är så små att de i praktiken är försumbara. Resultatet av studien visar således att metoderna är ungefär likvärdiga, dvs resultaten styrker Mc Donalds slutsats att diskriminantanalys inte nödvändigtvis behöver förkastas till förmån för logitanalys vid studier av val mellan två alternativ.

Mot diskriminantanalys talar dock dess striktare statistiska förutsättningar. Även om det faktum att dessa förutsättningar inte var uppfyllda i vårt fall inte tycks ha haft någon väsentlig inverkan på resultaten, är det naturligtvis inte uteslutet att det kan påverka resultaten i andra fall.

De utförda modellstudierna i bilaga 1 visar också att även en olämpligt utformad modell som är oanvändbar för prognosändamål kan återge färdmedelsandelarna nära nog exakt i det observationsmaterial som använts för estimeringen av modellen. Förmåga att återge färdmedelsandelar i observationsmaterialet är således ett otillräckligt krav på en modell och resultaten understryker vikten av att modellerna verkligen konstruerats på ett ur beteendeteoretisk synpunkt riktigt sätt.

6 REFERENSER

1. Lindahl, L & Eklind, B, 1972, Val mellan bil och kollektiva färdmedel i Norrköping (Statens råd för byggnadsforskning) R R 19:1972, Stockholm.

2. Hur parkeringsanläggningars utnyttjande beror på gångavstånd, parkeringsavgift och kollektiva resmöjligheter, 1974 (Allmänna Ingenjörsbyrån AB) Stockholm.

3. Berglund, Tegner, Widlert, 1977, Val av resmål och färdsätt vid inköpsresor - en beteendestudie (Statens råd för byggnadsforskning) Rapport R 8:1977, Stockholm.

4. de Donnea F X, 1971, The Determinants of transport mode choice in Dutch cities (Rotterdam University press) Rotterdam.

5. Warner, S L, 1962, Stochastic choice of mode in urban travel: A study in binary choice (Northwestern University Press)

6. Quarmby, D A, 1967, Choice of travel mode for the journey to work. Journal of Transport Economics an Policy, vol I, nr 3, sept. 1967.

7. Watson, P L, 1974, Choice of estimation procefure for models of binary choice. Regional and Urban Economics, nr 4, 1964.

8. Mc Donald, J F, Choice of estimation procedure for models of binary choice. Further Evidence.

9. Ben Akiva, M, A disaggregate direct demand model for simultaneous choice of mode and destination (International conference on Transportation Research) Belgien

10. Domencich A, Mc Fadden D, 1975, Urban Travel Demand (North-Holland publishing Company) New York.

BILAGA 1 METODJÄMFÖRELSE PÅ OBSERVATIONSMATERIAL FRÅN NORRKÖPING

I ansökan till detta forskningsprojekt avsågs att basera metodjämförelsen på ett datamaterial som samlats in 1968 i Norrköping av Allmänna Ingenjörsbyrån AB. Materialet från undersökningen kompletterades på vissa punkter och användes i ett forskningsprojekt där färdmedelsvalsmodeller togs fram med hjälp av diskriminantanalys (1). Avsikten var att göra en logitanalys på samma datamaterial. Logitanalysen skulle först göras med samma förklaringsvariabler som använts vid diskriminantanalysen. Eftersom diskriminantanalysen givit i olika avseende dåliga resultat skulle därefter logitanalysen göras med de förklaringsvariabler som använts vid den logitanalys AIB gjort på datamaterialet från Uppsala och Västerås (2) för att därigenom undersöka om mer rimliga förklaringsvariabler kunde ge bättre resultat. Dessa logitanalyser har också utförts och resultaten presenteras fortsättningsvis i denna bilaga. Vi kommer också att gå närmare in på orsakerna till att resultaten endast redovisas i bilageform. Datamaterialets struktur, i form av medelvärden, standardavvikelser, minimum och maximum för olika variabler, beskrivs närmare i bilaga 3. För övriga uppgifter om variabeldefinitioner, datainsamling etc hänvisas till de aktuella rapporterna.

I kapitel 2 beskrivs logit- och diskriminantanalysen närmare. För diskriminantanalysen som beskrivs i denna bilaga användes dock en något enklare linjär sannolikhetsmodell av formen:

$$P_b = L(X)$$

där P_b = sannolikheten att välja färdsättet bil

$L(X)$ = en linjär funktion av förklaringsvariabler

Logitmodellerna har fortfarande formen:

$$P_b = \frac{e^{L(X)}}{1+e^{L(X)}}$$

Eftersom modellerna har olika form är de erhållna koefficienternas siffervärden inte jämförbara mellan metoderna.

I tabellen nedan visas de erhållna koefficienterna från diskriminantmodellerna i rapport (1), samt motsvarande koefficientvärden för logitmodellerna med samma förklaringsvariabler. Vid diskriminantanalysen valdes förklaringsvariablerna ut med rent statistiska kriterier i ett program för s k stegvis diskriminantanalys.

Variabel X_j		Koefficient β_j (t-värde)			
		Modell 1 Diskri-minant	Logit	Modell 2 Diskri-minant	Logit
ln reslängd km		0.12 (2.40)	0.56 (1.59)	0.10 (2.00)	0.47 (1.30)
buss-spårvagn	1=buss 0=spårvagn	0.11 (1.83)	0.83 (2.06)	0.10 (1.67)	0.82 (1.94)
ln spilltids-kvot	koll./bil			0.06 (1.50)	0.37 (1.09)
lunchresa till bostaden	1=ja 0=nej			0.11 (1.83)	1.55 (2.03)
ln hushållsinkomst efter skatt				0.03 (1.50)	0.09 (1.65)
konstant		0.65 -	0.54 (1.25)	0.36 -	-0.83 (1.09)
X^2		-	7.39	-	16.83
R^2 respektive pseudo R^2		0.0529	0.0610	0.0784	0.1356

Siffrorna inom parantes är t-värden som utgör ett test på om koefficienten är signifikant skild ifrån noll (se kapitel 3). En jämförelse mellan de erhållna t-värdena visar att båda metoderna ger ungefär likvärdiga resultat i detta avseende. Signifikanserna är ungefär lika stora, och de skillnader som finns tycks variera slumpmässigt mellan metoderna. t-värdena är låga. I modell 1 är ln reslängd signifikant på 95%-nivån i diskriminantmodellen och buss-spårvagn i logitmodellen. I modell 2 är fortfarande ln reslängd enda signifikanta variabel på 95%-nivån i diskriminantmodellen. I logitmodellen är endast variabeln lunchresa signifikant. Ingen variabel i någon modell är signifikant på 99%-nivån. Det erhållna X^2-värdet visar att logitmodellerna i sig är signifikanta på 95 respektive 99%-nivån.

Samtliga koefficienter har förväntade tecken, exempelvis ökar sannolikheten att välja bil när reslängden ökar, vilket är rimligt.

Multipla korrelationskoefficienterna (R) visar att logitmodellerna ger något bättre förklaringsgrad än diskriminantmodellerna. Korrelationskoefficienterna är dock mycket låga.

De statistiska testen visar således att modellerna har mycket låg förklaringsgrad och låga signifikanser för ingående koefficienter. De valda förklaringsvariablerna förefaller inte heller lämpliga. Att, som i modell 1, försöka förklara färdmedelsvalet med enbart reslängden och en variabel för om det är buss eller spårvagn som är alternativ till bilen, är knappast realistiskt.

I tabellen nedan visas exempel på resultat som erhölls när nya logitmodeller estimerades med i huvudsak samma förklaringsvariabler som användes i rapport (2) (logitanalysen av arbetsresor i Uppsala och Västerås). Dessa förklaringsvariabler är alla mer beteendeteoretiskt välmotiverade än de tidigare använda. Eftersom variablerna inte var definierade på samma sätt, så görs ingen direkt jämförelse av de erhållna modellerna från de två undersökningarna.

Variabel X_j	Koefficient β_j (t-värde)		
	Modell 3	Modell 4	Modell 5
Reskostnadsdifferens (kollektivkostnad minus rörliga kostnader bil)	-0.440 (1.14)	-0.791 (2.04)	-0.663 (1.80)
Restidsdifferens (tid i fordon, kollektivt minus bil)	-0.011 (0.84)		
Gångtidsdifferens (tid utanför fordon, kollektivt minus bil)	0.001 (0.05)		
Bytestid kollektivt	0.099 (1.61)		
Väntetid kollektivt	-0.098 (0.97)		
Spilltid bil		-0.057 (1.38)	
Totaltidsdifferens (dörr-till-dörr, kollektivt minus bil)			-0.004 (0.32)
Totaltidsdifferens exklusive bilspilltid		-0.008 (0.62)	
Konstant	2.077 (2.36)	2.180 (4.12)	1.594 (4.41)
χ^2	8.46	6.45	4.23
pseudo R^2	0.0711	0.0545	0.0359

χ^2-värdena visar att modell 3 och 5 endast är signifikanta på 80%-nivån, och modell 4 på 90%-nivån. Även pseudo R^2-värdena är mycket låga.

I modell 3 är ingen variabel signifikant på 95%-nivån och kostnadsdifferensen, tidsdifferensen samt väntetid kollektivt har fel tecken. I modell 4 är kostnadsdifferensen signifikant men samtliga variabler i modell 4 och 5 har fel tecken. I alla modellerna är tidsdifferens- och kostnadsdifferensvariablerna starkt korrelerade vilket är en orsak till de dåliga resultaten.

Vi har således inte kunnat påvisa att någon av de angivna variablerna har inverkan på färdmedelsvalet. Eftersom vi från beteendeteori, andra undersökningar och från "sunt förnuft" vet att tidsskillnader, kostnadsskillnader, väntetider etc faktiskt inverkar på färdmedelsvalet så är resultatet synnerligen dåligt. Vi återkommer senare till de tänkbara orsakerna till detta.

För att testa modellernas klassificeringsförmåga sattes varje individs variabelvärden in i respektive modell. Om P_b då blev större än 0.5 klassificerades individen som biläkare, annars som kollektiväkare. Nedan visas antalet felklassificerade separat för de som verkligen åkt med bil respektive åkt kollektivt. Totalt åkte i verkligheten 169 personer bil och 33 personer kollektivt.

Modell	Antal felklassificerade biläkare	Antal felklassificerade kollektiväkare
1 diskriminant	0	33
logit	0	33
2 diskriminant	1	32
logit	0	32
3 logit	0	33

Praktiskt taget samtliga individer i observationsmaterialet klassificerades som biläkare, oavsett vilket färdsätt de i verkligheten använt. Det kan för övrigt konstateras att diskriminantmodell 1 aldrig kan klassificera en individ som kollektiväkare. Den lägsta sannolikheten som kan erhållas är nämligen 0.65 (=konstanten).

Eftersom modellen beskriver en individs sannolikhet att välja bil och modellerna enligt vad som visats ovan givit sannolikheter större än 0.5 för praktiskt taget samtliga som i verkligheten åker kollektivt så är det helt uppenbart att modellerna inte alls beskriver individernas val av färdmedel på ett acceptabelt sätt.

Om vi beräknar sannolikheten att välja bil för respektive person i observationsmaterialet, därefter summerar sannolikheterna och dividerar summan med totala antalet observationer, så erhålls den beräknade bilandelen. Nedan jämförs den beräknade bilandelen för varje modell med den verkliga:

		Bilandel procent
Verklig bilandel i observationsmaterialet		83.7
modell 1	diskriminant	85.3
	logit	83.7
modell 2	diskriminant	91.1
	logit	83.7
modell 3	logit	82.8

Logitmodellerna ger i stort sett helt rätt bilandel medan diskriminantmodellerna ger en något för hög bilandel. Resultatet kan skenbart förefalla mycket bra, särskilt för logitmodellerna. Av den tidigare redovisade klassificeringstesten framgick att modellerna inte alls var acceptabla. Denna senare beräkning visar bara att modellestimeringen lyckats så till vida att den totala bilandelen återges ungefär riktigt för observationsmaterialet. Som påpekats i kapitel 3 innebär detta inte att modellerna är användbara för prognosändamål. För detta krävs att modellerna verkligen beskriver och förklarar trafikanternas beteende.

Vi kan sammanfattningsvis konstatera att modellresultaten är synnerligen dåliga. Modellkoefficienternas signifikanser är mycket svaga och modellernas klassificeringsförmåga är obefintlig. De flesta förklaringsvariabler som vi a priori vet har stor inverkan på färdmedelsvalet får vid modellestimeringen fel tecken, eller blir inte signifikant skilda från noll. De skillnader som ändå kan upptäckas talar till logitanalysens fördel. Den multipla korrelationskoefficienten är högre och bilandelen återges bättre. Eftersom modellerna är så dåliga är det dock mycket vanskligt att dra några säkra slutsatser av skillnaderna.

I fortsättningen av denna bilaga skall vi försöka närmare förklara orsakerna till det erhållna resultatet.

En viktig orsak har vi redan berört, nämligen samvariationen mellan tids- och kostnadsvariabler. Variabler som samvarierar påverkar skattningen av varandras koefficienter, dvs när två variabler samvarierar kan man inte säkert identifiera hur mycket av färdmedelsvalet som beror på vardera av variablerna (det s k multikollinearitetsproblemet). Detta är ett mycket svårlösligt problem eftersom t ex de rörliga bilkostnaderna alltid kommer att samvariera med restid för bil.

Själva datamaterialet är relativt litet, 202 observationer. Detta är särskilt problematiskt eftersom endast 16% åkt kollektivt (33 personer). Empiriska erfarenheter har visat att sample under ca 200 observationer ger stora medelfel i koefficientberäkningen, men då har det varit sample med jämnare storleksfördelning mellan de två grupperna.

Datamaterialet saknar uppgifter om användning av bil under arbetstiden, en faktor som i andra undersökningar har visat sig spela mycket stor roll för färdmedelsvalet vid arbetsresor.

I enkäten som användes för att samla in datamaterialet frågades efter den totala restiden från dörr till dörr, samt hur stor del av denna som var spilltid (tid utanför fordonet). Vid enkätbearbetningen delades därefter spilltiden upp i komponenter efter vissa schablonregler. Därigenom blev variationen i dessa variabler liten vilket leder till svårigheter att identifera deras betydelse.

En översiktlig granskning av det kodade materialet samt av de ifyllda enkäterna har också medfört vissa tvivel på själva grundmaterialets kvalitet. Det är dock svårt att så lång tid efter undersökningens genomförande (nio år) dra några säkra slutsatser.

Slutligen kan det också vara på sin plats att säga något om principerna för variabelurval i den genomförda diskriminantanalysen. Urvalet skedde så att estimeringsprogrammet plockade ut förklaringsvariabler i den ordning de ur rent statistisk synpunkt gav störst bidrag till modellens förklaringsgrad. Antalet förklaringsvariabler bestämdes i sin tur genom test av om den sist medtagna variabeln var signifikant, och om den gav "signifikant bidrag" till modellens förklaringsgrad. Detta är ett principiellt sett mycket tveksamt tillvägagångssätt. Om man har tillräckligt många - i och för sig oväsentliga - variabler tillgängliga är det ändå troligt att någon av dem skenbart tycks bidra till den statistiska "förklaringen" av färdmedelsvalet. Estimeringsprogrammet kan självfallet inte veta om det är ett verkligt orsakssamband som spårats eller bara ett skenbart samband. Felaktigt medtagna variabler kan sedan i sin tur "slå ihjäl" de variabler som borde varit med. Problemställningen kan ytterligare belysas genom två citat från Moshe Ben-Akiva (9):

"The specification of a travel demand model necessarily embodies some assumptions about the relationships among the variables underlying travel behaviour. Predictions made by the model are conditional on the correctness of the behavioral assumptions and, therefore, are no more valid than the behavioral assumptions on which the model is based."

"In general, it is impossible to determine the correct specification of a model from data analysis. It should be determined from theory or a priori knowledge which are based on experience with, and understanding of, the phenomenon to be modelled."

---

(9) Ben-Akiva, M, A disaggregate direct demand model for simultaneous choice of mode and destination (International conference on Transportation Research) Belgien.

BILAGA 2 Medelvärden, standardavvikelser, minima och maxima för datamaterialet från Uppsala och Västerås.

bil = gruppen bilåkare, alt = gruppen som åkt med något alternativt färdsätt, tot = samtliga

Variabel		Medelvärde	Standard-avvikelse	Minimum	Maximum
gångtid bil, parkering till arbete, min per dag	bil	6.99	3.75	4.00	16.00
	alt	8.78	4.99	4.00	36.00
	tot	7.74	4.39	4.00	36.00
bil i tjänsten minst en gång per vecka, 1 = ja, 0 = nej	bil	0.41	0.49	0.00	1.00
	alt	0.14	0.35	0.00	1.00
	tot	0.30	0.46	0.00	1.00
kostnadsdifferens (kostnad alt färdsätt minus parkeringskostnad bil) kronor per dag	bil	0.84	2.24	-8.00	14.50
	alt	-0.20	2.27	-8.00	6.60
	tot	0.41	2.31	-8.00	14.50
tidsdifferens exkl gångtid bil, alt färdsätt minus bil. Minuter per dag	bil	31.52	21.12	-8.00	144.00
	alt	24.00	19.79	-18.00	84.00
	tot	28.38	20.89	-18.00	144.00

BILAGA 3 Medelvärden, standardavvikelser, minima och maxima för datamaterialet från Norrköping.

bil = gruppen bilåkare, koll = gruppen kollektivåkare, tot = samtliga

Variabel		Medelvärde	Standard-avvikelse	Minimum	Maximum
ln reslängd km	bil	1.05	0.52	- 0.69	2.30
(enkelt avstånd)	koll	0.86	0.57	- 0.92	1.87
	tot	1.02	0.53	- 0.92	2.30
buss-spårvagn					
1 = buss, 0 = spårvagn	bil	0.77	0.42	0.00	1.00
	koll	0.58	0.50	0.00	1.00
	tot	0.74	0.44	0.00	1.00
ln spilltidskvot	bil	1.23	0.59	- 0.33	2.89
koll/bil	koll	1.03	0.67	- 0.18	2.53
	tot	1.20	0.61	- 0.33	2.89
lunchresa till bostad	bil	0.22	0.42	0.00	1.00
1 = ja, 0 = nej	koll	0.06	0.24	0.00	1.00
	tot	0.20	0.40	0.00	1.00
ln hushållsinkomst	bil	9.55	2.52	- 4.61	11.04
efter skatt, kr	koll	8.67	4.28	- 4.61	10.92
	tot	9.41	2.88	- 4.61	11.04
reskostnadsdifferens	bil	-0.28	0.75	- 3.45	1.00
(kollektivkostnad minus	koll	-0.03	0.60	- 1.81	1.04
rörliga bilkostnader)	tot	-0.24	0.74	- 3.45	1.04
kronor per dag					
restidsdifferens (tid i	bil	26.52	17.63	- 8.00	90.00
fordon, koll minus bil)	koll	24.30	21.21	- 6.00	94.00
minuter per dag	tot	26.16	18.23	- 8.00	94.00
gångtidsdifferens (tid	bil	4.27	7.63	-20.00	34.00
utanför fordon, koll	koll	3.27	9.19	-10.00	40.00
minus bil)	tot	4.11	7.89	-20.00	40.00
bytestid kollektivt	bil	4.21	4.51	0.00	20.00
minuter per dag	koll	2.48	3.28	0.00	10.00
	tot	3.93	4.38	0.00	20.00
väntetid kollektivt	bil	7.59	1.94	6.00	10.00
minuter per dag	koll	7.94	2.03	6.00	10.00
	tot	7.64	1.95	6.00	10.00
spilltid bil	bil	7.34	4.35	2.00	36.00
minuter per dag	koll	8.24	4.21	4.00	20.00
	tot	7.49	4.33	2.00	36.00
totaltidsdifferens	bil	33.86	17.77	6.00	100.00
exklusive bilspilltid	koll	32.55	20.85	2.00	106.00
minuter per dag	tot				

BILAGA 4 Logitmodellen

I denna bilaga skall vi relativt kortfattat beskriva den ekonomiska teori som leder fram till logitmodellen (och vissa andra modeller). En utförlig behandling av denna teori finns i t ex "Urban Travel Demand" av D. Mc Fadden och T. Domencich (10). Avsnittet bygger huvudsakligen på denna referens.

Modeller som beskriver konsumenters beteende bygger på att individen handlar rationellt, att han kan rangordna tänkbara alternativ i angelägenhetsordning och att han alltid väljer det alternativ som han finner mest önskvärt med hänsyn till sina individuella preferenser. Valet sker inom de ramar som ges av individens tillgängliga tid och inkomst. Konsumenten försöker således maximera sin nytta inom de tillgängliga resursramarna.

I vanlig ekonomisk teori tänks konsumenten efterfråga en viss mängd av en viss vara eller nyttighet. Individens efterfrågefunktion är kontinuerlig, exempelvis leder en marginell prisförändring till en marginell efterfrågeförändring. Mot varje pris svarar således en viss bestämd efterfrågan. Denna teori kan inte direkt tillämpas för att studera efterfrågan inom trafikområdet. Individens efterfrågan inom detta område kännetecknas nämligen i allmänhet av att den är diskret till sin natur, inte kontinuerlig. Om vi exempelvis betraktar valet av färdsätt så leder en prisförändring på ett färdsätt antingen till att individen byter färdsätt eller också till att han fortsätter att använda samma färdsätt som tidigare. En marginell prisförändring för ett färdsätt leder således på individnivå inte till en marginell efterfrågeförändring. Istället för att beskriva hur en viss efterfrågan kontinuerligt förändras kommer vi därför att behandla ett val mellan ett ändligt antal ömsesidigt uteslutande handlingsalternativ.

De val som framför allt är aktuella när vi studerar trafik är valet av:

- bostad och arbetsplats
- bilinnehav
- resfrekvens för olika ändamål
- destination för olika resor
- tidpunkt på dagen för olika resor
- färdsätt
- färdväg

De två första punkterna illustrerar mer långsiktiga val som individen gör, medan de övriga sker med ett kortare tidsperspektiv. Olika val kan tänkas ske samtidigt (simultant), eller i en viss ordning (sekvensiellt). Denna fråga som i sig är mycket väsentlig behandlas inte i föreliggande rapport, utan den intresserade hänvisas till exempelvis referens (3) och (10). Rapporten behandlar modeller för val av färdmedel

---

(10) Domencich A, Mc Fadden D, 1975, Urban Travel Demand (North-Holland publishing Company) New York

vid arbetsresor (där ju frekvens, tidpunkt och destination i allmänhet är givna). De i denna bilaga härledda modellerna är dock helt generella och kan användas för samtliga val-situationer och för olika sekvensiella eller simultana beslutsstrukturer.

Låt oss anta att en viss individ har J olika alternativ att välja mellan. Vi betecknar alternativen j = 1, 2, 3, ..., J. Varje alternativ kan vara t ex en resa med ett visst färdsätt, en resa till ett visst färdmål, en resa med ett visst färdsätt till ett visst färdmål osv. Olika individer kan ha olika alternativ (och olika antal alternativ) att välja mellan. Varje alternativ j = 1, 2, ... J, som individen kan välja emellan har en vektor av observerade egenskaper X^J (t ex restider och reskostnader för ett visst färdsätt). Individens observerade socioekonomiska egenskaper betecknas med vektorn S (t ex ålder, kön och utbildning). Vi antar att individen har en nyttofunktion som mäter individens nytta av varje alternativ. Nyttan för ett visst alternativ antas vara en funktion av alternativets egenskaper X, individens socioekonomiska egenskaper S och av en oobserverad vektor ε som innehåller alla de egenskaper hos alternativet och alla de karakteristika för individen som vi inte kunnat observera och mäta (t ex personlig smak och erfarenhet). Nyttofunktionen kan då skrivas:

$$u = U(X, S, \varepsilon)$$

Om individerna i vår trafikundersökning väljs ut slumpmässigt från en delpopulation av individer med gemensamma socioekonomiska karakteristika S och gemensamma alternativ, så blir vektorn ε stokastisk och därigenom blir även värdena på nyttofunktionen stokastiska. För att förenkla beteckningarna kan vi då skriva

$$u = U(X, S) \qquad (1)$$

där u är en stokastisk variabel vars värde beror på exakt vilken individ vi har dragit från delpopulationen med samma observerade karakteristika och alternativ.

Individen väljer ett visst alternativ i om detta är det alternativ som maximerar hans nytta, dvs individen väljer alternativ i om

$$U(X^i, S) > U(X^j, S) \quad \text{för } j \neq i,\ j = 1, \ldots, J \qquad (2)$$

Eftersom värdena på nyttofunktionen är stokastiska så inträffar händelsen i ovanstående ekvation med en viss sannolikhet

$$P_i = P\left[U(X^i, S) > U(X^j, S) \quad \text{för } j \neq i,\ j = 1, \ldots, J\right] \qquad (3)$$

Den stokastiska nyttofunktionen U (X, S) kan utan förlust av generalitet skrivas

$$U(X, S) = V(X, S) + \eta(X, S) \qquad (4)$$

där V inte är stokastisk utan avspeglar populationens "representativa värdering", medan η är stokastisk och avspeglar individuella olikheter samt skillnader i icke observerade egenskaper för alternativen. Ekvation (3) kan då skrivas

$$P_i = P\left[V(X^i, S) + \eta(X^i, S) > V(X^j, S) + \eta(X^j, S) \text{ för } j \neq i,\ j = 1, \ldots, J\right] \quad (5)$$

Formen på funktionen V och variabeln η:s fördelningsfunktion påverkas dels av teorin för individers beteende, dels av rent beräkningstekniska synpunkter.

Vi antar i detta sammanhang att V har formen

$$V(X, S) = Z^1(X, S)\,\beta_1 + \ldots\ldots + Z^k(X, S)\,\beta_k = Z(X, S)\,\beta \quad (6)$$

där $Z^k(X, S)$ är en empirisk funktion utan okända parametrar och $\beta = (\beta_1, \ldots, \beta_k)$ är en vektor av okända parametrar. Detta gör V till en linjär funktion av parametervektorn β vilket avsevärt underlättar estimeringen. Variablerna $Z^1, \ldots, Z^k$ kan vara olika transformationer av rådata (t ex logaritmer, differenser, kvoter) och de kan även vara kombinationer av socioekonomiska karakteristika och egenskaper hos alternativ. Den valda formen för V är därigenom fullt tillräckligt generell för våra ändamål.

Olika fördelningar för η_1 och η_2 ger upphov till olika modeller. Om de exempelvis är simultant normalfördelade erhålles den s k probitmodellen som för n=2 får formen

$$P\left[V_1 + \eta_1 \geq V_2 + \eta_2\right] = \Phi\left(V(X, S) - V(X^2, S)\right)$$

Den s k logitmodellen erhålles om η_i har en Weibullfördelning (extremvärdesfördelning, Gnedenkofördelning) dvs om

$$P\left[\eta_i \leq \eta\right] = e^{-e^{-(\eta + \alpha)}}$$

där α är en parameter. Man kan visa (se referens (10)) att om η_i - variablerna har oberoende Weibullfördelningar med parametrar α_i för i = 1, ..., n så gäller för n=2 att

$$P\left[V_1 + \eta_1 \geq V_2 + \eta_2\right] = \frac{e^{V_1 - \alpha_1}}{e^{V_1 - \alpha_1} + e^{V_2 - \alpha_2}} \quad (7)$$

där $V_j = V(X^j, S)$, samt allmänt att

$$P\left[V_1 + \eta_1 \geq V_i + \eta_i \text{ för } i = 2, \ldots, n\right] = \frac{e^{V_1 - \alpha_i}}{\sum_{i=1}^{n} e^{V_i - \alpha_i}} \qquad (8)$$

Parametern α_i kan absorberas in i definitionen av $V(X^i, S)$ genom att låta vissa Z^i i ekvation (6) vara lika med 0 för alla alternativ utom ett. Därigenom kan samtliga α_i sättas till noll utan att någon generalitet förloras. Ekvation (7) och (8) kan då skrivas

$$P_1 = \frac{e^{V(X^1, S)}}{e^{V(X^1, S)} + e^{V(X^2, S)}}$$

Val mellan två alternativ (binärt val)

respektive

$$P_i = \frac{e^{V(X^i, S)}}{\sum_{j=1}^{J} e^{V(X^j, S)}}$$

Val mellan flera alternativ

BILAGA 5 Transformering av diskriminantfunktion samt beräkning av noggrannhetsmått.

Diskriminantfunktionen z beräknades med ett standardprogram (BMD 04 M, Biomedical computer programs, UCLA). Transformeringen av z följer den beräkningsgång som angivits av Quarmby (6). Där finns även en teoretisk motivering till transformeringen. Här beskrivs endast beräkningsgången i korthet.

Diskriminantfunktion z skall transformeras till en sannolikhetsmodell av formen:

$$P_b = \frac{\frac{n_2}{n_1} \cdot e^{k \cdot z + c}}{1 + \frac{n_2}{n_1} e^{k \cdot z + c}} \qquad \text{(eller allmänt } P_b = \frac{e^{L(Z)}}{1 + e^{L(Z)}}\text{)}$$

där P_b = sannolikheten att välja bil
n_2 = antalet bilåkare
n_1 = antalet som ej åkt bil
k, c = konstanter
z = diskriminantfunktionen
L(Z) = funktion av z

Multiplikationen med n_2/n_1 motiveras av att om antalet som använder respektive färdsätt inte är lika så finns det en förhandssannolikhet (skild från 0.5) för att det ena färdsättet skall väljas oftare än det andra.

Konstanterna k och c kan bestämmas med linjär regression. Om gruppernas fördelningar standardiseras genom att t ex bilåkarnas fördelning reduceras med n_1/n_2 så kan regressionen göras på ett uttryck av formen:

$$\log (P_b / (1 - P_b)) = k \cdot z + c$$

där P_b är den standardiserade andelen bilåkare i ett visst intervall av z. Observationerna rangordnas först efter z-värde och indelas i lika breda klasser. Regressionsanalysen görs med $\log (P_b / (1 - P_b))$ som beroende variabel och medelvärdet för z i intervallet som oberoende variabel. I vårt fall gav detta följande resultat:

$$\log (P_b / (1 - P_b)) = 461.39 \cdot z - 0.0314$$

korrelationskoefficient = 0.96

antal klasser = 20

$$\text{dvs} \quad P_b = \frac{e^{461.39 \cdot z - 0.0314}}{1 + e^{461.39 \cdot z - 0.0314}}$$

eller i icke standardiserad form

$$P_b = \frac{e^{461.39 \cdot z - 0.0314 + 0.3337}}{1 + e^{461.39 \cdot z - 0.0314 + 0.337}}$$

där $0.3337 = \log n_2/n_1$

Signifikanserna för koefficienterna i diskriminantfunktionen kan bestämmas genom att dividera koefficientvärdet med medelfelet i koefficientberäkningen (ger t-värdet). Begreppet "medelfel i koefficienter" har ingen direkt tolkning för diskriminantfunktionen, men kan härledas genom diskriminantkoefficienternas proportionalitet mot motsvarande koefficienter från multipel regressionsanalys. Även t-värdesberäkningen följer i huvudsak den beräkningsgång som angivits av Quarmby (där de nedan angivna formlerna härleds). Skillnaden består av en skalfaktor $n_1 + n_2 - 2$ som kommer in eftersom det standardprogram vi använt ger en diskriminantfunktion som är beroende av samplestorleken.

$$t_k = \lambda_k \,/\, d\,(\lambda_k)$$

där t_k = t-värdet för koefficient k

λ = diskriminantfunktionskoefficienten

$d\,(\lambda_k)$ = medelfelet för λ

$$d\,(\lambda_k) = \frac{1 + K \cdot a}{K} \sqrt{s^2 \cdot c_{kk}^{-1}} \,/(n_1 + n_2 - 2)$$

$$\text{där} \quad K = \frac{n_1 \cdot n_2}{(n_1 + n_2)\,(n_1 + n_2 - 2)}$$

n_2 = antalet bilåkare

n_1 = antalet icke bilåkare

$$a = (n_1 + n_2 - 2)\,(\bar{Z}_1 - \bar{Z}_2) =$$

$$= (n_1 + n_2 - 2) \cdot \sum_{k=1}^{m} \lambda_k \,(\bar{X}_{k1} - \bar{X}_{k2})$$

m = antalet variabler

$$S^2 = \frac{n_1 \cdot n_2}{n_1 + n_2} \left(1 - \frac{k \cdot a}{1 + k \cdot a}\right) / (n_1 + n_2 - m - 1)$$

C^{-1}_{kk} är ett element i den kvadratsummematris som skulle använts i en regressionsanalys, och inte den kvadratsummematris som används i diskriminantanalysen. Skillnaden är att matrisen i regressionsanalysen beräknas över samtliga observationer, medan matrisen vid diskriminantanalysen beräknas genom att addera de två gruppernas kvadratsummematriser.

Korrelationskoefficienten R ges av:

$$R^2 = \frac{k \cdot a}{1 + k \cdot a}$$

SAMMANFATTNING

Inledning

Föreliggande undersökning syftar till att empiriskt jämföra två metoder för att analysera färdmedelsval, logitanalys och diskriminantanalys. Jämförelsen görs genom att använda båda metoderna på samma datamaterial och därigenom studera hur väl de kan förklara det observerade färdmedelsvalet.

Analysmetoderna

De båda metoderna används i rapporten för att estimera modeller som uttrycker sannolikheten för att välja ett visst färdsätt. Modellerna har samma matematiska form i båda fallen, men bygger på olika antaganden och förutsättningar. De har formen:

$$P_b = \frac{e^{L(X)}}{1 + e^{L(X)}}$$

där P_b = sannolikheten att en viss individ väljer bil som färdmedel

$L(X)$ = en linjär funktion av variabler som förklarar färdmedelsvalet

Funktionen $L(X)$ kan skrivas

$$L(X) = \beta_o + \sum_{j=1}^{n} \beta_j X_j$$

där X_j = variabler som förklarar färdmedelsvalet

β_o = konstant som beror på de faktorer som ej tagits med i analysen

β_j = koefficienter för de olika variablerna

Koefficienterna i modellerna bestäms med statistiska metoder så att de anpassar sig så väl som möjligt till ett givet datamaterial. I vårt fall används ett datamaterial från Uppsala och Västerås bestående av 496 observationer. Datamaterialet är insamlat genom arbetsplatsenkäter och avser valet mellan bil och "bästa" alternativa färdmedel vid arbetsresor gjorda av individer med reell tillgång till bil.

Jämförelse mellan analysmetoderna

De båda metoderna användes för att estimera modeller med ovan visat utseende. Det visade sig att samma förklaringsvariabler gav bäst resultat med båda metoderna samt att logitmodellen gav något högre förklaringsgrad än diskriminantmodellen men att skillnaden var liten. Tidsvärdena erhållna ur de båda modelltyperna blev mycket lika.

Diskriminantanalysen gav aningen bättre resultat när modellernas prognosförmåga testades, medan logitmodellen bättre återgav bilandelen i hela materialet och dessutom klassificerade individerna något bättre.

Slutsats

Resultatet av den empiriska jämförelsen visar att logitmodellen ger bättre resultat i vissa test och att diskriminantmodellen ger bättre resultat i andra. Vid en praktisk användning av modellerna är skillnaden så liten att de är helt likvärdiga.

Denna rapport hänför sig till forskningsanslag 750594-6 från Statens råd för byggnadsforskning till Allmänna Ingenjörsbyrån AB, Stockholm

R13:1978

ISBN 91-540-2811-6
Statens råd för byggnadsforskning, Stockholm

Art.nr: 6600713

Abonnemangsgrupp:
Ingår ej i abonnemang

Distribution:
Svensk Byggtjänst, Box 1403
111 84 Stockholm

Cirkapris: 20 kr exkl moms

R13:1978 Diskriminant- och logitanalys — en metodjämförelse C-O Berglund, G Tegner, S Widlert