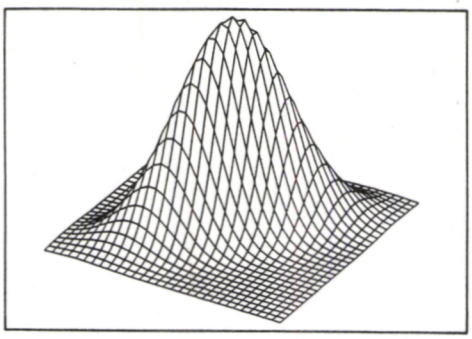
METSÄNTUTKIMUSLAITOKSEN
TIEDONANTOJA 279
Matemaattinen osasto

LOGISTISET JA LOGLINEAARISET MALLIT JA NIIDEN RATKAISEMINEN BMDP-OHJELMISTOLLA

Risto Häkkinen ja Kimmo Linnilä



Helsinki 1987

VAPK Kampin VALTIMO
Helsinki 1987

METSÄNTUTKIMUSLAITOS
Jaloetusosasto

Metsäntutkimuslaitoksen tiedonantoja 279

LOGISTISET JA LOGLINEAARISET MALLIT JA NIIDEN RATKAISEMINEN BMDP-OHJELMISTOLLA

Risto Häkkinen ja Kimmo Linnilä

Helsinki 1987

SISÄLLYSLUETTELO

ISBN 951-40-0831-6
ISSN 0358-4283

Valtion painatuskeskus 1987

1. JOHDANTO JA ESIMERKKEJÄ

Tässä esityksessä tarkastellaan yhtäältä aineistoja, joiden havainnot on generoinut binäärinen prosessi: tapahtumalla on 2 tilaa, "onnistuminen tai epäonnistuminen", "oikea tai väärä", "elävä tai kuollut" jne. Toisaalta tarkastellaan aineistoja, joiden havainnot voidaan esittää tiivistetysti frekvensseinä kontingenssitaulukossa.

Tällaisia havaintoaineistoja on perinteisesti analysoitu menetelmillä, joista tunnetuimpia ovat suhteellisten osuuksien vertailutestit, χ^2-yhteensopivuustestit ja -riippumattomuustestit. Tässä esityksessä tarkastellaan toista menetelmäryhmää, logistisia ja loglineaarisia malleja ajatellen lähinnä Metsäntutkimuslaitoksen tutkimusaineistoja sekä laitoksella käytettävissä olevia tilastollisen tietojenkäsittelyn mahdollisuuksia. Näissä menetelmissä keskeisellä sijalla on ilmiöiden tilastomatemaattinen mallittaminen, jonka avulla riippuvuussuhteita voidaan analysoida perinteisiä menetelmiä syvällisemmin.

Seuraava esitys koostuu teoriaosasta (luvut 1 ja 2) sekä BMDP-ohjelman tulkituista tulostusesimerkeistä (liitteet 1 ja 2).

1.1. Logistinen malli

Oletetaan, että satunnaismuuttujalla Y on kaksi mahdollista tilaa, jotka koodataan "1" ja "0". Näiden tilojen todennäköisyydet ovat

$$P(Y=1) = \theta \; ; \; P(Y=0) = 1-\theta \; ; \; 0 \leq \theta \leq 1.$$

Tällaisen muuttujan sanotaan noudattavan Bernoullin jakaumaa, jonka odotusarvo ja varianssi ovat:

$$E(Y) = \theta,$$

$$D^2(Y) = \theta(1-\theta).$$

Bernoulli-jakautuneesta muuttujasta voidaan johtaa lukuisia muita jakaumia, mm. toistokokeista tutut binomijakauma, negatiivinen binomijakauma sekä geometrinen jakauma.

Logistisessa mallissa muuttujaa Y käsitellään "vastauksena" (response) erilaisiin "ärsykkeisiin" (stimulus). Terminologia tulee mallien sovelluksista lääketieteessä ja eläinkokeissa. Y on siis eräänlainen selitettävä muuttuja, jonka odotusarvoa tarkastellaan erilaisia (jatkuvia ja diskreettejä) taustamuuttujia vastaan. Asetelma muistuttaa lineaaristen mallien käytöstä tuttua tilannetta. Oleellinen ero on kuitenkin se, että selitettävä Y-muuttuja ei ole intervalliasteikolla mitattu jatkuvatyyppinen muuttuja, vaan se saa pelkästään arvoja 0 ja 1.

Esimerkki 1. 2x2- kontingenssitaulut

2x2- kontingenssitaulujen lähtökohtana on 2 ryhmää yksilöitä (tilastoyksiköitä), joista jonkin ominaisuuden suhteen on mitattu binäärisen muuttujan arvo. Tilastoyksikkönä voi olla esimerkiksi uudistusalan taimi, ryhmittelyn perusteena 2 erilaista maankäsittelymenetelmää ja binäärimuuttujana ominaisuus elossa/kuollut.

Molemmista ryhmistä poimitaan (esim. yhtä suuret) otokset, joiden koolla sinänsä ei ole mitään yhteyttä ryhmän edustaman perusjoukon suuruuteen. Otoksien antama informaatio voidaan tiivistää taulukkoon, jossa on kuvattu havaittujen tilastoyksikköjen lukumäärien jakaumat eri ryhmissä mitattavan ominaisuuden suhteen.

	Ryhmä1	Ryhmä2	Yht.
Onnistuneet	s_1	s_2	s_1+s_2
Epäonnistuneet	n_1-s_1	n_2-s_2	$n_1+n_2-s_1-s_2$
Yht.	n_1	n_2	n_1+n_2

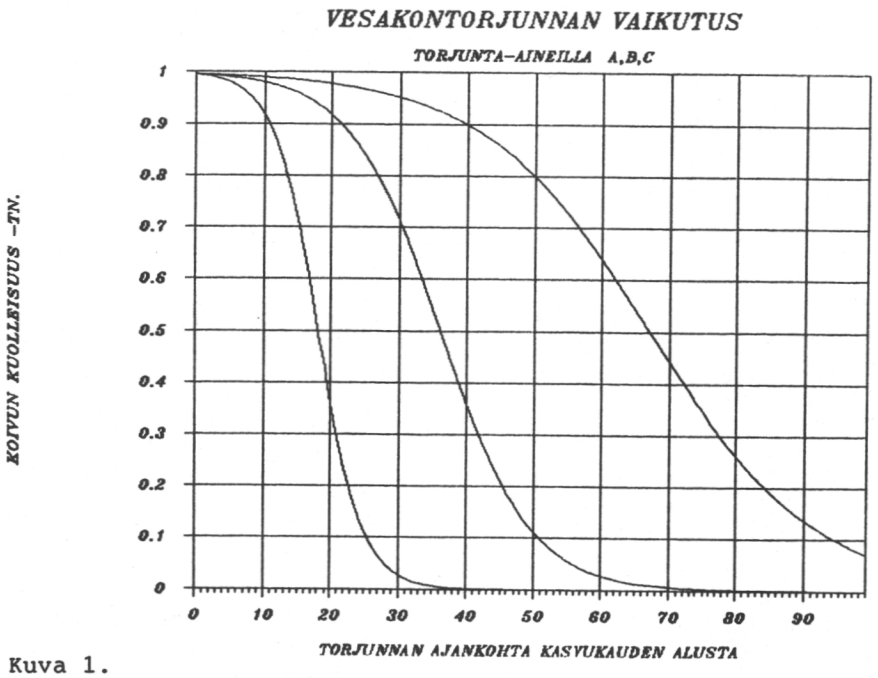
s_1/n_1 ja s_2/n_2 ovat tuntemattomien todennäköisyyksien θ_1 ja θ_2 tyhjentävät estimaattorit.

Kun testaamalla pyritään selvittämään, poikkeavatko ryhmien onnistumistodennäköisyydet toisistaan, oletetaan vain ryhmän vaikuttavan todennäköisyyteen. Näin ei asiantila todellisuudessa kuitenkaan välttämättä ole, vaan taustalla saattaa

olla tuntemattomia tai kontrolloimattomia tekijöitä. Ne aiheuttavat sen, ettei onnistumistodennäköisyys ryhmän sisällä olekaan vakio, jolloin joudutaan käyttämään mutkikkaampia malleja.

Esimerkki 2. Ärsyke-vastaus-riippuvuus.

Tehdään koe, jossa tarkastellaan vesakon kemiallista torjuntaa eri ajankohtina kasvukauden aikana männyn uudistusaloilla. Havaittavat suureet ovat lehtipuun eloonjääminen suhteellisina osuuksina ja ruiskutuksen ajankohta eri koealoilla. Mittausaineisto voidaan kuvata graafina, jossa x-akselina on aika (kalenteriaika, terminen kasvukausi) tai lämpösumma ja y-akselina eloonjääneiden koivunvesojen osuus koealoilla.



Kuva 1.

Esimerkissä binäärisen muuttujan käyttäytyminen riippuu ärsykemuuuttujasta ja sen odotusarvo vaihtelee systemaattisesti ärsykemuuttujan arvojen muuttuessa. Eräs ilmiön tarkasteluun liittyvä ongelma on sopivan matemaattisen funktion löytäminen riippuvuuden kuvaamiseksi.

1.2. Loglineaarinen malli

Frekvenssitaulukoita analysoitaessa ollaan useimmiten kiinnostuneita taulukointimuuttujien välisistä suhteista. Yksi- tai useampiulotteisen kontingenssitaulukon solufrekvenssi F ei ole luonteeltaan varianssianalyysin tapainen jatkuvatyyppinen selitettävä muuttuja, vaan lukumäärälaskuri, joka kuvaa nominaaliasteikolla mitatun monimuuttuja-aineiston jakaumaa. Loglineaarisessa mallissa taulukon solufrekvenssi esitetään taulukointimuuttujien ja niiden yhdysvaikutusten funktiona, joka muodollisesti muistuttaa varianssianalyysimallia. Kuitenkaan loglineaarisella mallilla ei varianssianalyysin tapaan testata frekvenssimuuttujan keskiarvoeroja, vaan ongelman asettelu on lähellä korrelaatio- ja faktorianalyysiä: mielenkiinnon kohteena on taulukointimuuttujien keskinäiset riippuvuudet ja vaikutukset.

Esimerkki 3. 3x2x2-kontingenssitaulu.

On haluttu tutkia kolmen taimikoissa esiintyvän tuhotyypin ja kahden taimilajin välisiä riippuvuuksia. Tutkimustaimikot on perustettu käyttäen kahta uudistamistapaa. Havaintoaineisto on kuvattu taulukkona, jonka solujen arvoina ovat taimista lasketut frekvenssit:

	Uudistamistapa 1		Uudistamistapa 2	
	Taimi-laji 1	Taimi-laji 2	Taimi-laji 1	Taimi-laji 2
Tuho 1	f_{111}	f_{112}	f_{121}	f_{122}
Tuho 2	f_{211}	f_{212}	f_{221}	f_{222}
Tuho 3	f_{311}	f_{312}	f_{321}	f_{322}

Logistinen malli ei tule kysymykseen taulukkoa analysoitaessa, koska aineistossa ei ole binääristä selitettävää muuttujaa. Perinteisillä χ^2-taulukkotesteillä voitaisiin käsitellä ainoastaan 3-ulotteisen taulukon 2-ulotteisia marginaalitaulukoita. Sen sijaan loglineaarisen mallin avulla voidaan taulukkomuuttujien välisiä riippuvuussuhteita tutkia syvällisemmin kuin χ^2-testeillä. Numeerinen esimerkki on esitetty liitteessä 2.

2. LOGISTISTEN JA LOGLINEAARISTEN MALLIEN TEOREETTISTA TAUSTAA

2.1. Yleisestä lineaarisesta mallista

Aluksi esitellään lineaarisen mallin teorian yleisiä tuloksia siinä laajuudessa ja esitysmuodossa, kuin se on tarpeellista itse aiheen käsittelemiseksi.

Tarkastellaan satunnaismuuttujia Y_i, joiden odotusarvon oletetaan noudattavan seuraavaa mallia:

$$E(Y_i) = \sum_{s=1}^{p} b_s x_{is} \quad , \text{ missä } \begin{matrix} i = 1,\ldots,n \\ s = 1,\ldots,p \end{matrix}$$

Matriisimerkinnöin asia voidaan ilmaista seuraavasti:

$E(Y) = Xb$, missä X on nxp-matriisi ja
b on px1-vektori:

$$X = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ & & \\ x_{n1} & \cdots & x_{np} \end{bmatrix} \qquad \text{ja} \qquad b = \begin{bmatrix} x_1 \\ \cdot \\ \cdot \\ \cdot \\ x_p \end{bmatrix} .$$

Tuntemattoman parametrivektorin b estimaattori $\hat{b}$ ratkaistaan pienimmän neliösumman menetelmällä minimoimalla lauseke $(Y-X\hat{b})'(Y-X\hat{b})$ $\hat{b}$:n suhteen. Tällöin saadaan:

$$\hat{b} = (X'X)^{-1}X'Y.$$

Mallin residuaalit ja jäännösneliösummat lasketaan kaavoilla:

$$\hat{R} = Y-\hat{Y} = Y-X\hat{b},$$

$$\hat{S}_R = \hat{R}'\hat{R} = (Y-\hat{Y})'(Y-\hat{Y}).$$

Yleisen lineaarisen mallin erikoistapauksia ovat regressio-, varianssi- ja kovariarianssimallit. Regressioanalyysissä

oletetaan momenttimatriisin olevan täyttä astetta, t.s. että matriisi X'X on epäsingulaarinen. Kovarianssi- ja varianssianalyysissä tämä ehto ei ole voimassa ja tällöin tuntemattomille parametreille joudutaan asettamaan lineaarisia reunaehtoja.

2.2 Pienimmän neliösumman menetelmä, kun selitettävä muuttuja on binäärinen

Oletetaan, että selitettävä muuttuja Y on binäärinen ja noudattaa seuraavaa lineaarista mallia:

$$\theta_i = P(Y_i=1) = E(Y_i) = \sum_{s=1}^{p} b_s x_{is} .$$

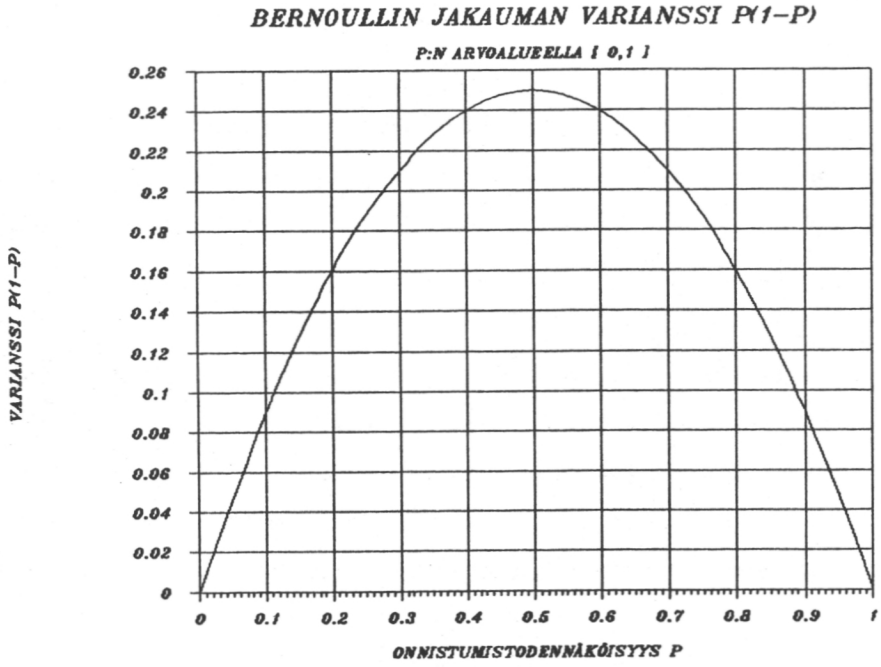
Mallin tuntemattomien parametrien estimointi voisi tapahtua pns-menetelmällä tavalliseen tapaan käsittelemällä Y:n arvoja 0 ja 1 kvantitatiivisina havaintoina. Mallin käytöllä on kuitenkin vakavia rajoituksia, koska Y on binäärinen.

Ensinnäkin θ:n arvoalue on välillä (0,1). Tämä rajoittaa mallin käyttöä, koska varsin helposti päästään selittävien muuttujien arvokombinaatioilla tilanteisiin, joissa θ:n "ennuste" on mieletön. Usein tämän ongelman kanssa joudutaan tekemisiin tosin silloinkin, kun muuttuja ei ole binäärinen.

Toinen ongelma aiheutuu muuttujien Y_i varianssien erisuuruudesta:

$$\mathrm{var}(Y_i) = \theta_i(1-\theta_i).$$

Y:n varianssi vaihtelee siis havainnosta toiseen. Tosin tiedetään, että kohtalaisetkaan poikkeamat variansseissa eivät vaikuta kohtuuttomasti estimaattoreiden tehokkuuteen. Jos θ:n arvot vaihtelevat välillä (0.2,0.8) on vastaava varianssin vaihteluväli (0.16,0.25).



Kuva 2.

Jos varianssin arvot vaihtelevat huomattavasti havainnosta toiseen, voidaan aluksi ratkaista yleinen lineaarinen malli ja laskea mallista kullekin havainnolle Y:n odotusarvon estimaatti, jota käytetään yllä olevassa varianssin kaavassa. Varianssiestimaattien käänteisarvoja käytetään painoina ja malli lasketaan uudestaan. Täysin tehokkaaseen estimointitulokseen ei joka tapauksessa päästä, koska käytetyt painot ovat mallista estimoituja. Koska Y_i:t eivät ole normaalisti jakautuneita, estimoitujen kertoimien merkitsevyyksiä ei voida testata eikä muodostaa luottamusvyöhykkeitä jne.

2.3. Logistinen malli

Seuraavassa esitetään selitettävän, binäärisen muuttujan Y odotusarvon $E(Y)=\theta$ ns. logistinen muunnos:

$$(1)\quad \theta_i = \frac{e^{x_i b}}{1 + e^{x_i b}}, \qquad i = 1,\ldots,n$$

$$(2)\quad 1-\theta_i = \frac{1}{1 + e^{x_i b}},$$

missä x_i on 1xp-rivivektori ja b px1-parametrivektori. Yksinkertaisella laskutoimituksella voidaan yhtälöt (1) ja (2) saattaa yhtäpitäväksi seuraavan muunnoksen kanssa:

$$\lambda_i = \log \frac{\theta_i}{1-\theta_i} = x_i b = \sum_{s=1}^{p} b_s x_{is}$$

eli $\Lambda = Xb$,

missä X on rivien x_i muodostama matriisi. Tässä muunnoksessa todennäköisyyden θ vedonlyöntisuhteesta $\theta/(1-\theta)$ on otettu logaritmi ja päädytty mallin linearisoimiseen. Menettelyä voidaan käyttää tilanteessa, jossa binäärinen aineisto on tiivistetty suhteellisiksi osuuksiksi (prosenteiksi). Esimerkissä 2 käsiteltiin tällaista tilannetta. Mikäli empiirisessä aineistossa ei ole θ:n arvoina nollia tai ykkösiä, voidaan käyttää tavallista lineaarista mallia. Mikäli taas selitettävä muuttuja on binäärinen, on turvauduttava muihin estimointimenetelmiin.

Tarkastellaan riippumattomia muuttujia $Y_1,\ldots,Y_n$ ja niiden reaalisaatioita $y_1,\ldots,y_n$ otoksessa. Otoksen likelihoodissa

$$L(y|X,b) = \prod_{i=1}^{n} P(y=y_i)$$ esiintyy tekijä

$$P(y_i=1) = \frac{e^{x_i b}}{1 + e^{x_i b}}$$,kun $y_i=1$ ja tekijä

$$P(y_i=0) = \frac{1}{1 + e^{x_i b}}$$,kun $y_i=0$.

Otoksen likelihood saa näin ollen muodon

$$(3) \quad L(y|X,b) = \prod_{i=1}^{n} \frac{e^{x_i b y_i}}{1 + e^{x_i b}} \quad .$$

Käytännössä logistisen mallin parametrit estimoidaan maximum likelihood -menetelmällä, t.s. määrittämällä parametrivektori b siten että otoslikelihood L(y|X,b) tulee mahdollisimman suureksi. Pienimmän neliösumman menetelmän käyttö ei tässä yhteydessä olisi mahdollistakaan. Ml-menetelmää käytettäessä tutkija uskoo valitsemansa mallin hyvyyteen ja pyrkii näin ollen maksimoimaan otoksen todennäköisyyden mallin puitteissa. Voidaan todistaa, että jos parametrilla on tehokas (efficient) estimaattori, ml-ratkaisu tuottaa sen. Mikäli lineaarisen mallin virhetermi noudattaa normaalijakaumaa, ovat pienimmän neliösumman ja ml-estimaattorit samat.

2.4. Loglineaarinen malli

Loglineaarisia malleja käytetään analysoitaessa moniulotteisten frekvenssitaulukoiden taulukointimuuttujien välisiä riippuvuussuhteita ja vaikutuksia. Loglineaarisella mallilla taulukon solufrekvenssien logaritmit kuvataan mallin parametrien lineaarisena funktiona. Parametrit edustavat taulukointimuuttujia ja niiden yhdysvaikutuksia. Kaksiulotteista taulukkoa kuvaava yksinkertainen loglineaarinen malli, kun rivi- ja sarakemuuttujat eivät riipu toisistaan, on muotoa

$$(4) \qquad \ln F_{ij} = m + A_i + B_j \text{ , missä}$$

$F_{ij} = E(f_{ij})$ = solufrekvenssin odotusarvo,
f_{ij} = havaittu solufrekvenssi,
m = vakio,
A_i = muuttujan A tason i vaikutusta kuvaava parametri,
B_j = muuttujan B tason j vaikutusta kuvaava parametri.

Malli (4) voidaan kirjoittaa myös tulomuotoisena

$$(5) \qquad F_{ij} = e^{m} e^{A_i} e^{B_j} .$$

Jos taulukon rivi- ja sarakemuuttujat riippuvat toisistaan, malliin (4) on lisättävä riippuvuutta kuvaava yhdysvaikutustermi

$$(6) \qquad \ln F_{ij} = m + A_i + B_j + (AB)_{ij} .$$

Johdannon esimerkissä 1 kuvattu 2-suuntainen taulukko voitaisiin logistisen mallin sijasta analysoida myös edellä esitettyjen loglineaaristen mallien avulla.

Mallit (4) ja (6) muistuttavat tuttua 2-suuntaista varianssianalyysimallia sillä erotuksella, että varianssianalyysin intervalliasteikolla mitatun selitettävän muuttujan sijasta on loglineaarisessa mallissa solufrekvenssin logaritmi. Kuitenkaan loglineaarista mallia käytettäessä ei olla kiinnostuneita varianssianalyysin tapaan frekvenssien keskiarvoeroista sinänsä, vaan taulukointimuuttujien välisistä riippuvuuksista ja vaikutuksista, joita frekvenssien keskinäiset suuruussuhteet heijastavat.

Mallilla (4) on myös kiinteä yhteys χ^2-riippumattomuustestistä tuttuun teoreettiseen (expected) solufrekvenssiin, joka perustuu nollahypoteesin mukaiseen oletukseen, että taulukon rivi- ja sarakemuuttujat eivät riipu toisistaan. Merkitään taulukon rivin i reunafrekvenssiä r_i :llä, sara-

kefrekvenssiä c_j :llä sekä havaintojen kokonaislukumäärää N:llä ja tulkitaan mallin (4) parametrit seuraavasti

$$A_i = \ln r_i ,$$
$$B_j = \ln c_j ,$$
$$m = -\ln N.$$

Sijoittamalla yllä olevat malliin (5) voidaan loglineaarinen malli (4) kirjoittaa muotoon

$$(7) \qquad F_{ij} = \frac{r_i \, c_j}{N} ,$$

mikä on täsmälleen sama kaava, jolla χ^2-testin teoreettiset solufrekvenssit lasketaan. Loglineaarisessa mallissa on näin saatu yhdistetyksi eräitä keskeisiä χ^2-taulukkotesteissä ja varianssianalyysissä käytettyjä periaatteita.

Taulukointimuuttujien riippuvuutta tutkittaessa verrataan loglineaarisia malleja (4) ja (6) toisiinsa testaamalla, poikkeaako mallin (6) yhdysvaikutustermi $(AB)_{ij}$ merkitsevästi nollasta. Kaksisuuntaisen taulukon ollessa kyseessä tämä vastaa χ^2-testin suorittamista.

Useampiulotteisen taulukon solufrekvenssin odotusarvo voidaan analogisesti kaavan (5) mukaisesti esittää tulomuotoisena mallina, jonka tekijät kuvaavat, varianssianalyysin käsittein lausuttuna, muuttujien pää- ja yhdysvaikutuksia

taulukon kussakin ruudussa. Ottamalla frekvenssistä logaritmi saadaan tulomuotoinen malli muunnetuksi parametrien lineaarikombinaatioksi, loglineaariseksi malliksi, jonka avulla taulukointimuuttujien välisiä riippuvuuksia voidaan testata. Esimerkiksi 3-suuntaisen taulukon (vrt. johdannon esimerkki 3 ja liite 2) loglineaarinen malli on

$$\ln F_{ijk} = m + A_i + B_j + C_k + (AB)_{ij} + (AC)_{ik} + (BC)_{jk} + (ABC)_{ijk}.$$

Loglineaaristen mallien estimoinnissa ja testaamisessa käytetään lineaaristen mallien ja maximum likelihood -estimoinnin teoriasta tunnettuja tilastotieteen menetelmiä. Kaikkia loglineaarisia malleja ei voida ratkaista analyyttisesti, mutta käyttäen iteratiivisia algoritmeja saadaan nykyisillä tietokoneohjelmilla mallien parametrien ml-estimaatit laskettuiksi. Edullisten teoreettisten ominaisuuksien ohella loglineaaristen mallien käytön tärkein peruste on kuitenkin se, että niiden avulla voidaan tarkastella samanaikaisesti usean muuttujan välisiä riipuvuussuhteita, kun taas χ^2-menetelmillä voidaan testata moniulotteisista taulukoista vain kaksiulotteisia marginaalitaulukoita.

2.5. Logistisiin ja loglineaarisiin malleihin liittyviä testejä

2.5.1. Mallin laatu

Estimoidun (logistisen tai loglineaarisen) mallin ja havaintoaineiston yhteensopivuutta (goodness of fit) testataan Pearsonin yhteensopivuuden χ^2-testisuureella

$$(8) \qquad \chi^2 = \sum \frac{(f_i - F_i)^2}{F_i}$$

tai vaihtoehtoisesti likelidood ratio -testisuureella

$$(9) \qquad G^2 = 2\sum f_i \log \frac{f_i}{F_i} \quad ,$$

missä F_i on kuhunkin taulukon ruutuun i liittyvä mallilla laskettu frekvenssi ja f_i havaittu frekvenssi. Molemmat testisuureet noudattavat asymptoottisesti χ^2-jakaumaa: mitä suurempi testisuureen arvo on, sitä huonompi on yhteensopivuus. Vapausasteina on taulukon ruutujen lukumäärän ja mallin riippumattomien parametrien lukumäärän erotus.

2.5.2. Muuttujien merkitsevyys

Haluttaessa testata, ovatko tietyt muuttujat tai niiden yhdysvaikutukset tutkitun ilmiön kannalta merkitseviä, ratkaistaan kaksi mallia: 1) täysi malli (full model) ja 2) rajoitettu malli (restricted model), jossa testattavat tekijät on poistettu mallista.

A. Logistisen mallin tekijöiden merkitsevyyttä tutkitaan täyden ja rajoitetun mallin likelihoodien (3) avulla. Likelihood ratio -testisuure

$$(10) \qquad 2 \log \frac{L_1}{L_2}$$

noudattaa asymptoottisesti χ^2-jakaumaa: mitä suurempi testisuureen arvo on, sitä enemmän testattavat tekijät selittävät ilmiötä. Vapausasteina on mallien riippumattomien parametrien lukumäärien erotus.

B. Loglineaarisen mallin tekijöiden merkitsevyyttä testattaessa käytetään hyväksi G^2-testisuureen (9) additiivisuusominaisuutta, jonka mukaan rajoitetun ja täyden mallin G^2-testisuureiden erotus

$$(11) \qquad G_2^2 - G_1^2$$

myös noudattaa asymptoottisesti χ^2-jakaumaa. (Vastaava ei päde χ^2-testisuureelle (8).) Erotusta (11) käytetään testisuureena, jonka arvo on sitä suurempi, mitä merkitsevämpiä testattavat tekijät ovat. Vapausasteina on G^2-testisuureiden vapausasteiden erotus.

2.6. Mallityypin valinta

Mallin valinta riippuu havaintoaineiston luonteesta eli havainnot generoineesta prosessista ja muuttujien mitta-asteikoista sekä tutkittavista hypoteeseista. Tämän esityksen mallittamisteemaan liittyen voidaan todeta seuraavat yleisperiaatteet:

1. Jos selitettävä muuttuja on intervalliasteikolla mitattu jatkuvatyyppinen muuttuja, päädytään koejärjestelystä ja selittävien muuttujien mitta-asteikoista riippuen lineaariseen tai epälineaariseen regressio-, varianssi- tai kovarianssianalyysiin (BMDP:n R- ja V-ohjelmat).

2. Jos havainnot ovat suhteellisia osuuksia (esim. kuolleisuusprosentti/koeala), käytetään selitettävänä muuttujana logistista muunnosta ja mallin parametrien estimoinnissa lineaarista regressio-, varianssi- tai kovarianssianalyysiä kuten kohdassa 1.

3. Jos selitettävä muuttuja on binäärimuuttuja ja jos jotkut selittävät muuttujat ovat jatkuvatyyppisiä muuttujia, käytetään frekvenssitaulukon analyysissä logistista mallia (BMDPLR). Jos mallissa ei ole jatkuvia muuttujia, voidaan käyttää myös loglineaarista mallia (BMDP4F).

4. Jos frekvenssitaulukossa ei ole binääristä selitettävää muuttujaa, käytetään taulukon analysoinnissa loglineaarista mallia (BMDP4F).

Eräät frekvenssiaineistot on siis mahdollista analysoida joko logistisella tai loglineaarisella mallilla. Kuitenkin mallin valinta on useimmiten selkeä. Jos ollaan kiinnostuneita dikotomisen, selitettävän muuttujan riippuvuudesta taustamuuttujista ja varsinkin jos joku tai jotkut taustamuuttujista ovat jatkuvatyyppisiä, käytetään logistista mallia. Jos taas talukointimuuttujien joukossa ei ole selkeästi selitettäväksi erottuvaa binääriseksi tulkittavissa olevaa muuttujaa, tutkitaan muuttujien riippuvuussuhteita loglineaarisella mallilla.

Liite 1

LOGISTINEN MALLI JA BMDPLR-OHJELMA, ESIMERKKI

Roustekohouman (sulamisen ja jäätymisen takia maan pintaan muodostuvan epätasaisen jääkerroksen aiheuttaman taimen kohoamisen) riippuvuutta eräistä taustamuuttujista tutkittiin kenttäkokeessa. Taimet istutettiin kesällä ja mitattiin syksyllä. Havaintoyksikkönä oli yksittäinen taimi. Selitettävä muuttuja roustekohouma on binäärinen: sitä esiintyy tai ei esiinny. Selittävistä muuttujista taimen pituus on jatkuvatyyppinen muuttuja ja luokkamuuttujia ovat taimilaji (kennotaimi/paljasjuurinen taimi) ja muokkaustapa (mätästys/mätästys + ojitus). Oletetaan, että kokeen tulokseksi on saatu taulukossa (12) esitetty frekvenssijakauma. Havainnollisuuden vuoksi on pituus jaettu neljän sentin luokkiin ja analyysissä on käytetty pituushavaintona luokkakeskusta.

(12)

Taimilaji	Muokkaus	Pituus(cm)	Roustekohouma		
			Ei	Kyllä	YHT.
Kenno-taimi	Mätästys	1-4	9	7	16
		5-8	54	27	81
		9-12	45	6	51
		13-16	2	0	2
		YHT.	110	40	150
	Mätästys + ojitus	1-4	2	10	12
		5-8	36	37	73
		9-12	13	13	26
		13-16	0	1	1
		YHT.	51	61	112
Paljas-juurinen taimi	Mätästys	1-4	8	0	8
		5-8	47	7	54
		9-12	72	3	75
		13-16	12	1	13
		YHT.	139	11	150
	Mätästys + ojitus	1-4	1	0	1
		5-8	33	14	47
		9-12	33	3	36
		13-16	5	1	6
		YHT.	72	18	90
		YHTEENSÄ	372	130	502

Koska selitettävä roustekohouma-muuttuja on binäärinen, ei aineistoa voida analysoida kovarianssianalyysillä. χ^2-testeillä voitaisiin 2-suuntaisia osataulukoita analysoimalla yrittää selvittää aineiston riippuvuuksia, mutta muuttujien lukuisuuden takia se on vaikeaa, ja joka tapauksessa pituutta ei voitaisi käsitellä jatkuvatyyppisenä, intervalliasteikon muuttujana vaan luokkamuuttujana, jolloin menetettäisiin osa sen sisältämästä informaatiosta. Toisaalta taulukko voitaisiin analysoida loglineaarisella mallilla, jolloin usean muuttujan välisiä riippuvuuksia saataisiin selvitetyksi, mutta silloinkin pituutta olisi pidettävä luokkamuuttujana.

Jos halutaan saada roustekohouman esiintymistodennäköisyydelle θ selitysmalli, jossa pituus on regressiotyyppinen selittävä muuttuja, on suositeltavaa käyttää logistista mallia

$$\theta = \frac{e^{\Sigma\, bx}}{1 + e^{\Sigma\, bx}} , \qquad (13)$$

missä mahdollisina x-muuttujina ovat taimilaji, muokkaustapa ja niiden yhdysvaikutus sekä pituus.

BMDPLR-ohjelma muodostaa logistisen mallin askeltaen. Ohjelma voidaan suorittaa joko eräajona tai interaktiivisena ajona, jolloin käyttäjä rakentaa mallin askel askeleelta itse harkiten, mitkä tekijät kussakin vaiheessa ovat mallissa mukana. Seuraavassa tulostusesimerkissä on käytetty eräajoa, jolloin ohjelma tiettyjä päättelysääntöjä käyttäen itsenäisesti suorittaa askeltavan logistisen analyysin.

BMDPLR (STEPWISE LOGISTIC REGRESSION) TULKITTU OSATULOSTUS

BMDPLR:ssä käytetään seuraavia merkintätapoja:

- Yhdysvaikutustekijä ilmaistaan *-merkin avulla ja muuttujien nimien sijasta voidaan käyttää niiden alkukirjaimia. Esimerkiksi muokkauksen ja taimilajin yhdysvaikutus merkitään joko MUOKKAUS*TAIMILAJI tai M*T.

- Logistinen malli kuvataan usein luettelemalla siinä olevat tekijät siten, että luokkamuuttujien yhdysvaikutustekijöihin sisältyviä alemman asteen tekijöitä ei erikseen ilmoiteta. Esimerkiksi malli, johon halutaan selittäjiksi taimilaji, muokkaustapa, niiden yhdysvaikutus, viljelykohta ja pituus, merkitään PITUUS,VILJELYKOHTA,TAIMILAJI*MUOKKAUSTAPA.

Tulosten tulkinnassa Ts(x) tarkoittaa testisuuretta, joka on esitetty kaavassa numero (x). Pienennetty, kehyksissä oleva teksti on tietokonetulostetta ja normaali teksti on tulosten tulkintaa.

Sivuilla 27-30 on esitetty varsinainen logistisen mallin askeltava laskenta ja sivuilla 31-34 tarkastellaan saadun mallin ominaisuuksia.

OHJAUSKÄSKYJONO:

```
(1) /INPUT     FILE=RIITTA.VARIABLE=4.FORMAT='(F3.0,F3.0,F4.1,F3.0)'.
    /VARIABLE  NAMES=MUOKKAUS,TAIMILAJI,PITUUS,ROUSTE.
(2) /REGRESS   DEPENDENT=ROUSTE.
               INTERVAL=PITUUS.
               CATEGORICAL=MUOKKAUS,TAIMILAJI.
               MODEL=PITUUS,MUOKKAUS*TAIMILAJI.
               START=OUT,OUT.
               MOVE=2,2.
               METHOD=MLR.
(3) /PRINT     CELLS=MODEL.
               HISTOGRAM.
               NEWS.
    /PLOT      XVAR=PREDPROB.
               YVAR=OBSPROP.
    /END
```

(1) Tiedoston luku ja muuttujien määrittelyt. Tiedostossa on kullakin rivillä yksi havainto, vrt. (4). LR-ohjelma osaa lukea myös taulukkomuotoista dataa.

(2) Analyysin määrittely. Selitettävän muuttujan DEPENDENT oletusarvot ovat 0 ja 1 (1=suotuisa tapahtuma, joka tässä esimerkissä tarkoittaa roustekohouman esiintymistä). Jatkuvatyyppiset selittävät muuttujat ilmoitetaan parametrilla INTERVAL ja luokkamuuttujat parametrilla CATEGORICAL. Parametrilla MODEL ilmaistaan analyysissä käytettävissä olevat tekijät. Yhdysvaikutustekijöihin sisältyviä alemman asteen tekijöitä ei tarvitse erikseen luetella. Esimerkkiajossa mahdollisina selittävinä muuttujina ovat PITUUS, MUOKKAUS, TAIMILAJI ja yhdysvaikutus M*T.

Askeltava analyysi aloitetaan mallista, jossa on mukana kaikki MODELissa määritellyt tekijät, jollei parametrillä START erikseen ilmoiteta, mitkä tekijät jätetään aluksi mallista pois. Esimerkissä lähdetään liikkeelle mallista, jossa ei ole yhtään tekijää. Parametrin START luettelo vastaa parametrin MODEL tekijäluetteloa. Jos yhdysvaikutus jätetään pois, myös siihen sisältyvät tekijät jäävät pois. Esimerkiksi jälkimmäinen OUT saa aikaan sen, että yhdysvaikutuksen M*T ohella myös tekijät MUOKKAUS ja TAIMILAJI jäävät pois.

Parametri MOVE määrittelee, kuinka monta kertaa tekijä voidaan askeltavassa analyyissä ottaa malliin tai poistaa siitä. Kun malli on määritelty parametrillä MODEL, oletus on nolla kertaa, muutoin kaksi kertaa. Luettelo vastaa MODEL-luetteloa.

METHOD=MLR tarkoittaa, että mallin tekijät valitaan kussakin askeleessa maximum likelihood ratio -menetelmällä. Oletusarvona on asymptoottinen menetelmä ACE, joka on nopeampi mutta ei niin tarkka kuin MLR. Lopullinen ajo kannattaa suorittaa MLR-menetelmällä.

(3) Tulostus- ja piirrosvaihtoehtojen määrittelyä.

NUMBER OF CASES TO BE PRINTED 10 (4)

BASED ON INPUT FORMAT SUPPLIED 1 RECORDS READ PER CASE.

CASE NO.	LABEL	1 TAIMILAJ	2 MUOKKAUS	3 PITUUS	4 ROUSTE
1		1	1	2.500	0
2		1	1	2.500	1
3		1	1	6.500	0
4		1	1	6.500	1
5		1	1	10.500	0
6		1	1	10.500	1
7		1	1	14.500	0
8		1	1	14.500	1
9		1	2	2.500	0
10		1	2	2.500	1

NUMBER OF CASES READ. 502

TOTAL NUMBER OF RESPONSES USED IN THE ANALYSIS 502.
SUCCESS . . . . . 130.
FAILURE . . . . . 372.

NUMBER OF DISTINCT COVARIATE PATTERNS . . . . . (16)

(4) Ohjelma tulostaa 10 ensimmäistä havaintoa, ellei ole toisin määrätty. Esimerkkitiedostossa on 502 havaintoa (havainto/rivi). Havaintoaineistossa on TAIMILAJIxMUOKKAUSxPITUUS=2x2x4=16 erilaista selittävien muuttujien arvokombinaatiota eli sama kuin taulukon (12) ruutujen lukumäärä. Pituusmuttujalla on esimerkissä vain neljä erilaista arvoa, mutta usein jatkuvalla muuttujalla on niitä huomattavasti enemmän, jolloin myös 'ruutujen' lukumäärä kasvaa.

```
 VARIABLE     VALUE OR          DESIGN VARIABLES
NO. N A M E  INTERVAL  FREQ     ( 1)

  2 MUOKKAUS       1     300      -1
                   2     202       1

  1 TAIMILAJ       1     262      -1
                   2     240       1
```

(5)

(5) Varianssianalyysin tapaan kutakin luokkamuuttujaa vastaa mallissa joukko apumuuttujia, jotka liittyvät luokkamuuttujien tasoihin. Design-muuttujien ja mallin parametrien avulla voidaan laskea mallilla ennustettuja arvoja, ks. (14).

/ PRINT NEWS.

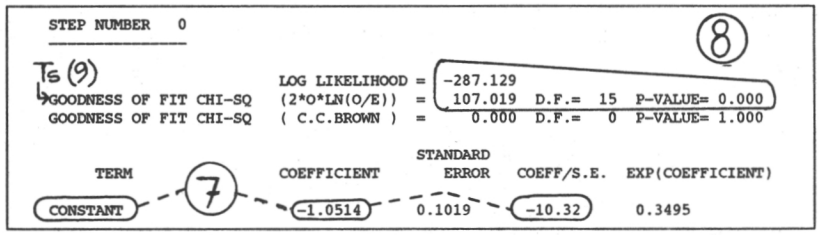
(6)

```
---------- N E W S ----------
Four columns are added to the summary report for diagnostic purposes:
  (1) 'CHI', an element of a goodness-of-fit Chi-square, used to detect
      Y outliers,
  (2) 'DEVIANCE', the element of a goodness-of-fit Chi-square,
      used to detect Y outliers,
  (3) 'HATDIAG', the diagonal element of the hat matrix, used to
      detect extreme cases in the X-space,
  (4) 'INFLUENCE', a measure similar to the Cook distance in
      linear regression.
   They are saved in a BMDP file if you state 'CONTENT=CELL.' .  These
results may also be plotted.

To plot cell results or variables state '/PLOT  XVAR = list.  YVAR = list.',
where list is a list of variable names and/or 'SUCCESS, FAILURE,
OBSPROP, PREDPROB, SEPRED, STDRESID, LOGODDS, CHI, DEVIANCE, HATDIAG,
or INFLUENCE'.

To print plots using the summary report with cells formed by variables
in the model one must specify 'CELLS=MODEL.' in the 'PRINT' paragraph.

The option word 'ALL' in the 'CELLS=' statement of the 'PRINT' paragraph has
been changed to 'USE'.  State 'CELLS=USE.' rather than 'CELLS=ALL.'.
-------------------------------------------------------------------------
   NEW FEATURES WHICH FIRST APPEARED IN THE 1985 RELEASE.

LINE EDITOR
  When BMDP programs are run interactively, the user frequently wants
  to correct and resubmit a problem.  A line editor has been added
  to all BMDP programs for the 1985 release making this easy to do.

TRANSFORMATION PARAGRAPH
--The symbol < may be used instead of 'LT' and the symbol > may be used
  instead of 'GT'.  In the conditional clause of the 'IF' statement 'IF(
  condition)THEN...', the equal sign (=) may be used in place of 'EQ'.
-------------------------------------------------------------------------
```

(6) Kuten kaikissa BMDP-ohjelmissa komennolla NEWS. saadaan tulostetuksi ohjelmiin tehdyt viimeisimmät muutokset. Ohjelmaan BMDPLR liittyvä uutisteksti on noin viisi kertaa niin pitkä kuin yllä oleva lyhennelmä, johon on koottu tärkeimmät viimeisen manuaalin ilmestymisen jälkeen tulleet uutuudet.

/ REGRESS

Tästä alkaa varsinaisen askeltavan logistisen analyysin tulostus (s. 27-30). Loppuosassa (s. 31-34) tarkastellaan saadun mallin ominaisuuksia.



(7) Nolla-askeleessa mallissa on vain vakio =-1.0514. Mallin parametrien merkitsevyys voidaan testata approksimatiivisen t-testisuureen avulla. Koska $|-10.32| >> 2$, vakio poikkeaa merkitsevästi nollasta. Malli on tässä vaiheessa

$$\theta = e^{-1.0514}/(1+e^{-1.0514}) = 0.2590.$$

(8) Kussakin askeleessa tulostetaan mallin likelihoodin (3) logaritmi, joka tässä tapauksessa on -287.129. Kun malli askel askeleelta paranee, sen likelihood kasvaa. Yhteensopivuuden χ^2-testisuure (9) =107.019 mittaa havaittujen ja mallilla laskettujen arvojen yhteensopivuutta. Koska $P \leq 0.000$, malli sopii erittäin huonosti havaintoihin, ks. 24 .

STATISTICS TO ENTER OR REMOVE TERMS

Ts (10)

TERM	APPROX. CHI-SQ. ENTER	D.F.	APPROX. CHI-SQ. REMOVE	D.F.	P-VALUE	LOG LIKELIHOOD
PITUUS	30.52	1			0.0000	-271.8699
MUOKKAUS	30.36	1			0.0000	-271.9509
TAIMILAJ	47.99	1			0.0000	-263.1326
M*T	IS OUT				MAY NOT BE ENTERED.	
CONSTANT			121.66	1	0.0000	-347.9599
CONSTANT			IS IN		MAY NOT BE REMOVED.	

(9)

(9) Jokaisen askeleen jälkeen ohjelma tutkii, onko syytä poistaa tai lisätä malliin joku tekijä. Kullekin vaihtoehdolle ohjelma laskee lisäys- tai poistotapahtuman mukaiselle uudelle mallille likelihoodin logaritmin ja maximum likelihood ratio-testisuureen (10) avulla tutkii, onko ko. tekijä merkitsevä vai ei. Esimerkissä tekijällä TAIMILAJI on suurin testisuureen (10) arvo

$$2(\log L_1 - \log L_2) = 2(-263.1326-(-287.129)) = 47.99,$$

joka on myös merkitsevin riskitasolla $P \leq 0.0000$. Seuraavassa askeleessa on siis lisättävä malliin TAIMILAJI. Ohjelma poistaa tekijän mallista, jos $P > 0.15$ (backward stepping) ja lisää tekijän malliin, jos $P \leq 0.10$ (forward stepping). Käyttäjä voi halutessaan määrätä uudet riskirajat. Jos menetelmäksi ei ole valittu METHOD=MLR (Maximum Likelihood Ratio), käytetään nopeampaa mutta epätarkempaa menetelmää ACE, joka johtaa edellä kuvatun χ^2-testin sijasta F-testisuureen käyttöön. Viimeisessä ajossa on varminta valita menetelmäksi MLR.

STEP NUMBER 1 TAIMILAJI IS ENTERED

Ts(10) (10)

LOG LIKELIHOOD = -263.133

IMPROVEMENT CHI-SQUARE (2*(LN(MLR)) = 47.993 D.F.= 1 P-VALUE= 0.000

GOODNESS OF FIT CHI-SQ (2*O*LN(O/E)) = 59.026 D.F.= 14 P-VALUE= 0.000

GOODNESS OF FIT CHI-SQ (C.C.BROWN) = 0.000 D.F.= 0 P-VALUE= 1.000

(11)

Ts(9)

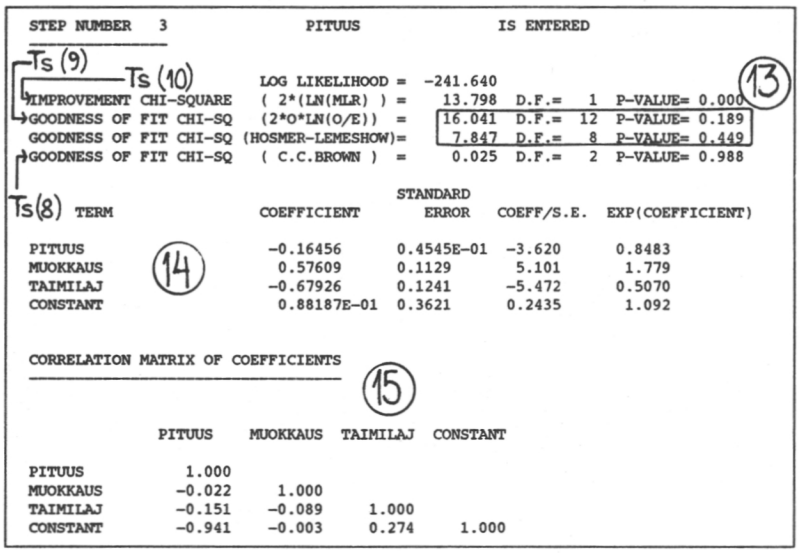
TERM	COEFFICIENT	STANDARD ERROR	COEFF/S.E.	EXP(COEFFICIENT)
TAIMILAJ	-0.75914	0.1176	-6.454	0.4681
CONSTANT	-1.2254	0.1176	-10.42	0.2936

(12)

(10) TAIMILAJIn lisääminen malliin paransi merkitsevästi mallin likelihoodia. Tämä selvisi myös kohdasta 9 .

(11) Havaintojen ja mallin yhteensopivuus ei vieläkään ole merkitsevä. Testiin on suhtauduttava kriittisesti, mikäli taulukon ruuduissa on vähän havaintoja kuten tässä esimerkissä, ks. (24).

(12) Mallin parametrit.



```
STEP NUMBER  3              PITUUS              IS ENTERED

                    LOG LIKELIHOOD =  -241.640
IMPROVEMENT CHI-SQUARE  ( 2*(LN(MLR) ) =    13.798  D.F.=   1  P-VALUE= 0.000
GOODNESS OF FIT CHI-SQ  (2*O*LN(O/E)) =    16.041  D.F.=  12  P-VALUE= 0.189
GOODNESS OF FIT CHI-SQ (HOSMER-LEMESHOW)=   7.847  D.F.=   8  P-VALUE= 0.449
GOODNESS OF FIT CHI-SQ  ( C.C.BROWN ) =     0.025  D.F.=   2  P-VALUE= 0.988

                                    STANDARD
TERM                 COEFFICIENT      ERROR    COEFF/S.E.   EXP(COEFFICIENT)

PITUUS               -0.16456      0.4545E-01  -3.620        0.8483
MUOKKAUS              0.57609      0.1129       5.101        1.779
TAIMILAJ             -0.67926      0.1241      -5.472        0.5070
CONSTANT              0.88187E-01  0.3621       0.2435       1.092

CORRELATION MATRIX OF COEFFICIENTS

             PITUUS    MUOKKAUS  TAIMILAJ  CONSTANT

PITUUS         1.000
MUOKKAUS      -0.022     1.000
TAIMILAJ      -0.151    -0.089     1.000
CONSTANT      -0.941    -0.003     0.274     1.000
```

(13) Askeleen 3 jälkeen mallilla laskettujen ja havaittujen frekvenssien yhteensopivuus on merkitsevä. Tässä esimerkkitapauksessa Ts(9) ei yksinään ole riittävän luotettava, koska kaikissa ruuduissa ei ole tarpeeksi havaintoja. Hosmer-Lemeshow-testi kuitenkin tukee vahvasti Ts(9):n antamaa tulosta. Katso (24) .

(14) Logistisen mallin (13) todennäköisyydet θ voidaan laskea mallin kertoimien ja design-muuttujien (5) avulla. Esimerkiksi, jos PITUUS=6.5, MUOKKAUS=1=mätästys, ja TAIMILAJI=2=paljasjuurinen, niin

$$\Sigma\, bx = 0.088187 - 0.16456*6.5 + 0.57609*(-1) - 0.67926*(1) = -2.2368$$

$\Sigma\, bx = \log(\theta/(1-\theta))$ = predicted log odd, ks (18).

Sijoittamalla edellä oleva kaavaan (13) saadaan mallilla laskettu roustekohouman todennäköisyys

$$\theta = e^{-2.2368}/(1+e^{-2.2368}) = 0.0965, \text{ ks. (18)} .$$

Mallin laskeminen on siis melko työlästä, mutta onneksi ohjelma laskee ennustetut todennäköisyydet kaikille aineistossa esiintyville arvokombinaatioille, ks. (18).

(15) Kunkin askeleen jälkeen ohjelma laskee mallin kertoimien välisen korrelaatiomatriisin. Pienet korrelaatiot merkitsevät kertoimien stabiiliutta ja selittäjien itsenäisyyttä.

STATISTICS TO ENTER OR REMOVE TERMS

TERM	APPROX. CHI-SQ. ENTER	D.F.	APPROX. CHI-SQ. REMOVE	D.F.	P-VALUE	LOG LIKELIHOOD
PITUUS			13.80	1	0.0002	-248.5390
MUOKKAUS			26.97	1	0.0000	-255.1245
TAIMILAJ			33.19	1	0.0000	-258.2335
M*T	0.00	1			0.9488	-241.6381
CONSTANT			0.06	1	0.8077	-241.6698
CONSTANT			IS IN			MAY NOT BE REMOVED.

(16)

NO TERM PASSES THE REMOVE AND ENTER LIMITS (0.1500 0.1000) .

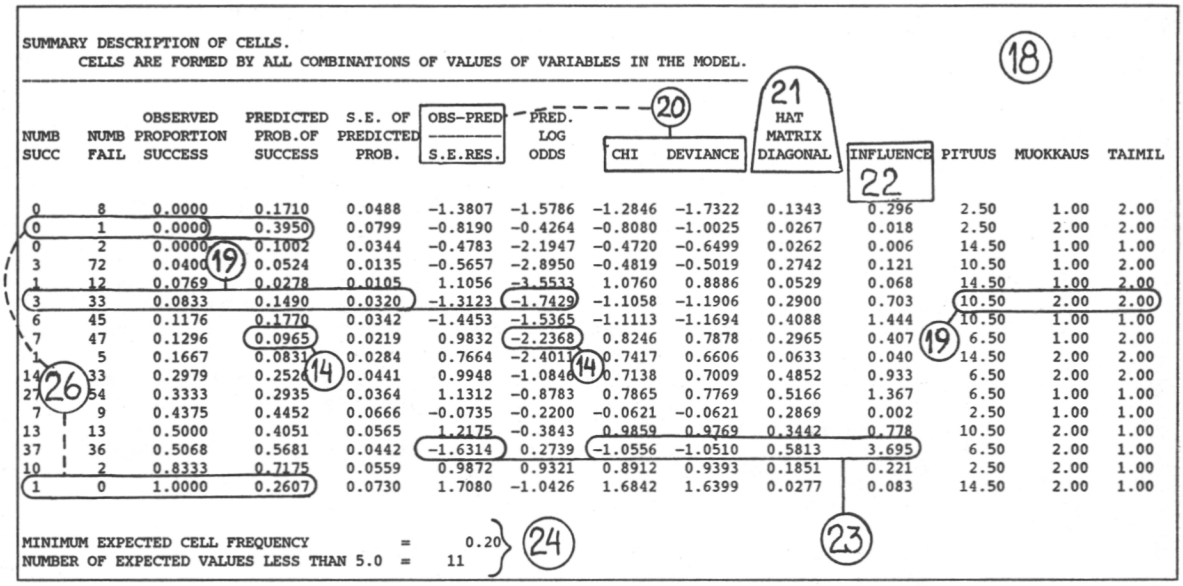
SUMMARY OF STEPWISE RESULTS

STEP NO.	TERM ENTERED REMOVED	DF	LOG LIKELIHOOD	IMPROVEMENT CHI-SQUARE	P-VAL	GOODNESS OF FIT CHI-SQUARE	P-VAL
0			-287.129			107.019	0.000
1	TAIMILAJ	1	-263.133	47.993	0.000	59.026	0.000
2	MUOKKAUS	1	-248.539	29.187	0.000	29.838	0.005
3	PITUUS	1	-241.640	13.798	0.000	16.041	0.189

(17)

(16) Askeleessa 3 tehty malli on lopullinen, yhtään tekijää ei voida poistaa mallista eikä yhdysvaikutustekijää M*T kannata lisätä malliin, koska se ei ole merkitsevä.

(17) Tiivistelmä analyysin vaiheista askeleittain



SUMMARY DESCRIPTION OF CELLS.
CELLS ARE FORMED BY ALL COMBINATIONS OF VALUES OF VARIABLES IN THE MODEL.

NUMB SUCC	NUMB FAIL	OBSERVED PROPORTION SUCCESS	PREDICTED PROB.OF SUCCESS	S.E. OF PREDICTED PROB.	OBS-PRED / S.E.RES.	PRED. LOG ODDS	CHI	DEVIANCE	HAT MATRIX DIAGONAL	INFLUENCE	PITUUS	MUOKKAUS	TAIMIL
0	8	0.0000	0.1710	0.0488	-1.3807	-1.5786	-1.2846	-1.7322	0.1343	0.296	2.50	1.00	2.00
0	1	0.0000	0.3950	0.0799	-0.8190	-0.4264	-0.8080	-1.0025	0.0267	0.018	2.50	2.00	2.00
0	2	0.0000	0.1002	0.0344	-0.4783	-2.1947	-0.4720	-0.6499	0.0262	0.006	14.50	1.00	1.00
3	72	0.0400	0.0524	0.0135	-0.5657	-2.8950	-0.4819	-0.5019	0.2742	0.121	10.50	1.00	2.00
1	12	0.0769	0.0278	0.0105	1.1056	-3.5533	1.0760	0.8886	0.0529	0.068	14.50	1.00	2.00
3	33	0.0833	0.1490	0.0320	-1.3123	-1.7429	-1.1058	-1.1906	0.2900	0.703	10.50	2.00	2.00
6	45	0.1176	0.1770	0.0342	-1.4453	-1.5365	-1.1113	-1.1694	0.4088	1.444	10.50	1.00	1.00
7	47	0.1296	0.0965	0.0219	0.9832	-2.2368	0.8246	0.7878	0.2965	0.407	6.50	1.00	2.00
1	5	0.1667	0.0831	0.0284	0.7664	-2.4011	0.7417	0.6606	0.0633	0.040	14.50	2.00	2.00
14	33	0.2979	0.2526	0.0441	0.9948	-1.0846	0.7138	0.7009	0.4852	0.933	6.50	2.00	2.00
27	54	0.3333	0.2935	0.0364	1.1312	-0.8783	0.7865	0.7769	0.5166	1.367	6.50	1.00	1.00
7	9	0.4375	0.4452	0.0666	-0.0735	-0.2200	-0.0621	-0.0621	0.2869	0.002	2.50	1.00	1.00
13	13	0.5000	0.4051	0.0565	1.2175	-0.3843	0.9859	0.9769	0.3442	0.778	10.50	2.00	1.00
37	36	0.5068	0.5681	0.0442	-1.6314	0.2739	-1.0556	-1.0510	0.5813	3.695	6.50	2.00	1.00
10	2	0.8333	0.7175	0.0559	0.9872	0.9321	0.8912	0.9393	0.1851	0.221	2.50	2.00	1.00
1	0	1.0000	0.2607	0.0730	1.7080	-1.0426	1.6842	1.6399	0.0277	0.083	14.50	2.00	1.00

MINIMUM EXPECTED CELL FREQUENCY = 0.20
NUMBER OF EXPECTED VALUES LESS THAN 5.0 = 11

(18) Taulukosssa on mallin muuttujien arvokombinaatioihin eli soluihin liittyviä tietoja, joiden avulla mallia ja sen ominaisuuksia voidaan tutkia. Solut ovat havaitun roustekohoumatodennäköisyyden mukaan nousevassa järjestyksessä (muukin järjestys on mahdollista). Esimerkissä näyttää pääsääntöisesti olevan pieniä todennäköisyyksiä, kun TAIMILAJI=2 ja MUOKKAUS=1, ja suuria kun TAIMILAJI=1 ja MUOKKAUS=2.

(19) Esimerkiksi solussa, jossa PITUUS=10.5, MUOKKAUS=2 ja TAIMILAJI=2, roustekohoumaa on 3 taimessa ja 33 taimessa ei ole, roustekohouman havaittu todenäköisyys on 0.0833 ja mallilla laskettu todennäköisyys on 0.1490 (keskivirhe 0.0320) ja vedonlyöntisuhteen logaritmi on -1.7429.

(20) Standardoitu residuaali (OBS-PRED)/S.E.RES, CHI ja DEVIANCE mittaavat, kuinka paljon malli ja havainnot ko. solussa poikkeavat toisistaan. Mitä lähempänä nollaa arvot ovat, sitä pienempää on poikkeama.

(21) Hattumatriisin diagonaalielementin suuruus kuvaa solun poikkeavuutta x-muuttujien suhteen. Mitä suurempi arvo on, sitä suurempi potentiaalinen vaikutus solulla on malliin.

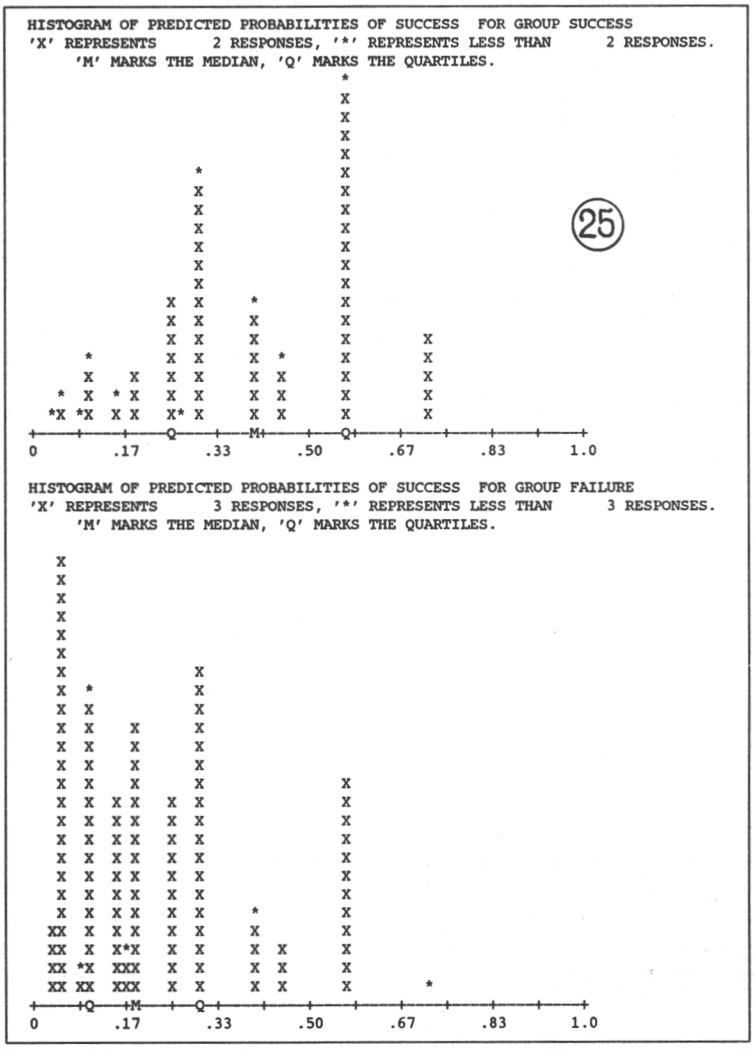
(22) INFLUENCE mittaa solun vaikutusta mallin kertoimiin eli sitä, kuinka voimakkaasti malli muuttuu, jos kyseinen havainto jätettäisiin laskuista pois. Mitä pienempi arvo on, sitä vähemmän vaikutusta solulla on mallin kertoimiin.

(23) Kohtien (20), (21) ja (23) perusteella havainto on jostain syystä poikkeava (tutkijan tulisi selvittää ja ymmärtää miksi). Taulukoista (12) ja (18) ilmenee, että kennotaimilla ylipäätään, kun muokkaus on mätästys+ojitus, on pituudesta riippumatta runsaasti roustekohoumia.

(24) Eräs peukalosääntöehto χ^2-testin pätevyydelle on, että yksikään odotettu frekvenssi ei saa olla alle yhden ja että korkeintaan 20 % frekvensseista saa olla alle viiden. Tässä esimerkissä χ^2-yhteensopivuustestin luotettavuus on suuntaa-antava, koska 11 ruudussa 32:sta on mallilla laskettu frekvenssi ≤ 5. Tosin pudottamalla yhden ja kahden havainnon ruudut pois jää jäljelle vain 5 huonoa ruutua, mutta niistäkin kahdessa on odotettu frekvenssi ≤ 1. Esimerkiksi ruudussa PITUUS=2.5, MUOKKAUS=2, TAIMILAJI=1 on ei-roustekohoumien (fail) odotettu frekvenssi =(1-0.7175)*(10+2)=3.4.

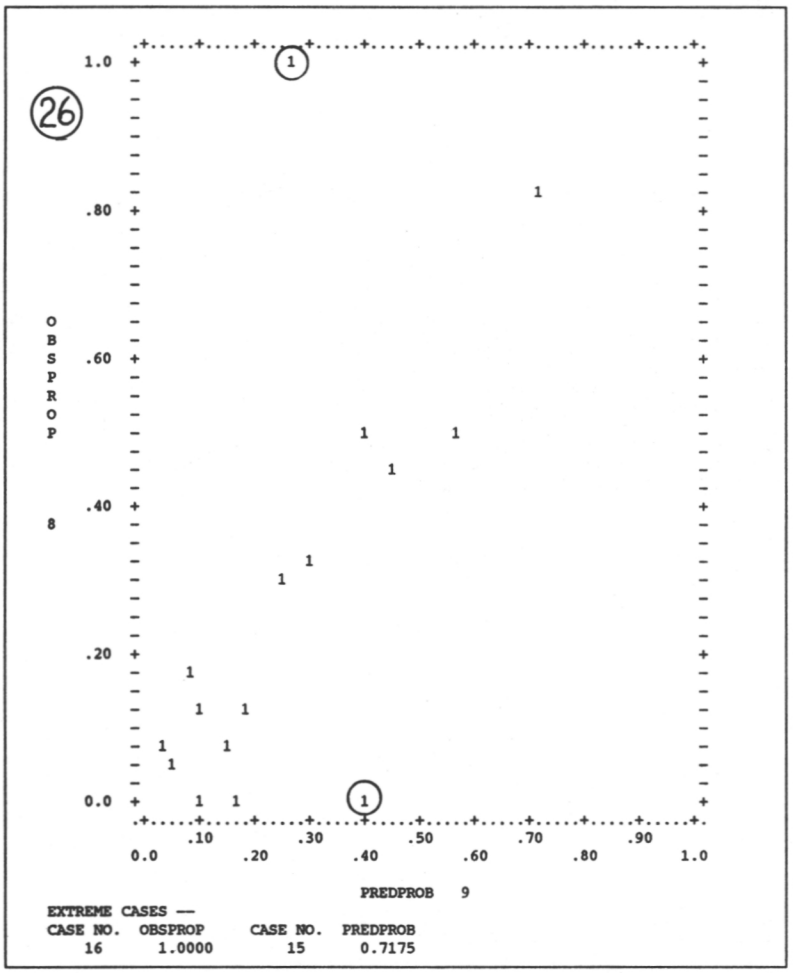
Hosmer-Lemeshow-testi on hyvä vaihtoehto tai tuki testisuureelle (9), kun havaintoja/ruutu on vähän. Siinä yhdistetään ruutuja siten, että mallilla laskettujen todennäköisyyksien vaihteluväli jaetaan 10 luokkaan siten, että niissä on suunnilleen sama määrä havaintoja. Näin saadusta taulukosta (2x10=20 ruutua) lasketaan χ^2-testisuure (8). Esimerkissä (ks. (13)) Hosmer-Lemeshow-testisuureen arvo on 7.847 ja $P \leq 0.449$ eli yhteensopivuus on jopa parempi kuin Ts(9):n mukaan. Hosmer-Lemeshow-testi ei ole mielekäs, jollei ruutuja ole paljon. (Esimerkiksi askeleisssa 0,1 ja 2, kun pituus ei ole mallissa, ruutuja on liian vähän.)

/ PRINT HISTOGRAM



(25) Yläkuva esittää niiden mallilla laskettujen roustekohouman todennäköisyyksien jakaumaa, jotka on saatu käyttämällä selittävien muuttujien arvoina niitä 130 havaintoa, joissa taimessa on ollut roustekohouma. Alakuvassa on käytetty niitä 372 havaintoa, joissa roustekohoumaa ei ole ollut. Kuvista voidaan päätellä mallin erottelu- tai ennustuskyvyn olevan kohtalainen: malli antaa roustekohoumaryhmässä suurempia roustekohouman todennäköisyyksiä kuin ryhmässä, jossa ilmiötä ei esiinny; osassa havaintoaineistoa mallin ennusteet menevät myös päällekkäin. Mallin erottelukyky on kuitenkin kärjekkäämpi kuin ensi silmäykseltä saattaisi luulla, sillä yläkuvassa risti edustaa 2 havaintoa kun taas alakuvassa 3 havaintoa.

/ PLOT XVAR=PREDPROB. YVAR=OBSPROP.



(26) Havaittujen ja mallilla laskettujen todennäköisyyksien ristikuvasta voi paljastua jotain mielenkiintoista. Kutakin aineistossa olevaa selittävien muuttujien arvokombinaatiota edustaa kuvassa yksi piste, tässä tapauksessa yhteensä 16 pistettä. Ihannetapauksessa pisteet asettuisivat vasemmasta alakulmasta oikeaan yläkulmaan kulkevalle suoralle. Kuvassa olevat kaksi poikkeavaa pistettä osoittautuvat kuitenkin harmittomiksi, sillä molemmat edustavat vain yhtä taimea, katso (18) .

Liite 2

LOGLINEAARINEN MALLI JA BMDP4F-OHJELMA, ESIMERKKI

Oletetaan, että johdannon esimerkissä 3 on saatu seuraavat mittaustulokset:

	Uudistamistapa 1		Uudistamistapa 2	
	Taimi-laji 1	Taimi-laji 2	Taimi-laji 1	Taimi-laji 2
Tuho 1	448	201	101	159
Tuho 2	252	109	259	431
Tuho 3	103	51	180	271

Eräs lähestymistapa tutkia taimilajin ja tuhon välistä riippuvuutta on testata χ^2-testillä kaksisuuntaista marginaalitaulukkoa (14), joka on saatu yhdistämällä uudistamistavat 1 ja 2:

(14)

	Taimilaji 1	Taimilaji 2
Tuho 1	549	360
Tuho 2	511	540
Tuho 3	283	322

Testin tulos on, että taimilajit ja tuhot riippuvat voimakkaasti toisistaan (χ^2= 36.99, D.F.=2, P≤0.00). - Jos kuitenkin testataan erikseen kummallekin uudistamistavalle taimilaji- ja tuhomuuttujan osataulukot, niin testin tulos on, että taimilajit ja tuhot eivät lainkaan riipu toisistaan (uudistamistapa 1: χ^2= 0.43, D.F.=2, P≤0.81; uudistamistapa 2: χ^2=0.66, D.F.=2, P≤0.72).

Testitulosten ristiriita selittyy sillä, että kolmisuuntaisen taulukon kaksisuuntaisesta marginaalitaulukosta (14) on kadonnut informaatio siitä, että uudistamistapa vaikuttaa oleellisesti tuhon ja taimilajin välisiin suhteisiin. Edellä esitetyn tapainen harhapäätelmä on usein mahdollinen analysoitaessa moniulotteisia taulukoita χ^2-testeillä, koska niillä voidaan tutkia vain 2-suuntaisia marginaalitaulukoita. Sen sijaan loglineaarista mallia käyttäen saadaan helposti selville, että taimilajin ja tuhon välinen yhdysvaikutus ei ole merkitsevä, kuten käy ilmi seuraavasta BMDP4F-ohjelman osatulostuksesta.

BMDP4F (LOGLINEAR MODEL) TULKITTU OSATULOSTUS

BMDP4F:ssä käytetään seuraavia merkintätapoja:

- Muuttujaan viitataan usein sen nimen alkukirjaimella, esimerkiksi T=TUHO, U=UUDTAPA ja L=LAJI.

- Loglineaarinen (hierarkkinen) malli kuvataan usein luettelemalla siinä olevat tekijät siten, että yhdysvaikutustekijöihin sisältyviä alemman asteen tekijöitä ei erikseen ilmoiteta. Esimerkiksi TU,LU tarkoittaa mallia F=m+T+U+L+TU+LU ja T,LU mallia F=m+T+U+L+LU.

Tulosten tulkinnassa Ts(x) tarkoittaa testisuuretta, joka on esitetty kaavassa numero (x). Pienennetty, kehyksissä oleva teksti on tietokonetulostetta ja normaali teksti on tulosten tulkintaa.

OHJAUSKÄSKYJONO:

```
(1) /INPUT    VARIABLE=4.FILE=TAULUKKO.FORM='(3F1.0,F4.0)'.
    /VARIABLE NAMES=TUHO,UUDTAPA,LAJI,FREKVENSSI.
    /CATEGORY NAMES(1)=TUHO1,TUHO2,TUHO3.
              NAMES(2)=UUDTAPA1,UUDTAPA2.
(2)           NAMES(3)=TAIMLAJ1,TAIMLAJ2.
    /TABLE    INDICES=LAJI,TUHO,UUDTAPA.
              COUNT=FREKVENSSI.
    /FIT      ASSOCIATION=3.
(3) /FIT      ALL.
    /FIT      MODEL=LTU.
              DELETE=SIMPLE.
              STEP=5.
    /PRINT    EXPECTED.
(4)           DIFFERENCES.
              LAMBDA.
    /END
```

(1) Taulukkodatan luku ja määrittely.

(2) Taulukon määrittely.

(3) Loglineaaristen mallien sovitus.

(4) Mallilla laskettujen frekvenssien ja mallin parametrien tulostus.

Tässä esimerkissä /FIT- ja /PRINT-kappaleet (paragraphs) on koottu samaan ohjauskäskyjonoon, mutta käytännössä loglineaaristen mallien sovitus kannattaa suorittaa erillisinä ajoina, /FIT-kappale kerrallaan. Yleensä vasta lopuksi on järkevää /PRINT-kappaleella tulostaa taulukon teoreettiset frekvenssit ja mallin parametrit, mikäli niitä tutkimalla halutaan tarkentaa käsitystä taulukointimuuttujien käyttäytymisestä.

/ TABLE INDICES=LAJI,TUHO,UUDTAPA.COUNT=FREKVENSSI.

UUDTAPA	TUHO	LAJI		
		TAIMLAJ1	TAIMLAJ2	TOTAL
UUDTAPA1	TUHO1	448	201 ö	649
	TUHO2	252	109 ö	361
	TUHO3	103	51 ö	154
	TOTAL	803	361 ö	1164
UUDTAPA2	TUHO1	101	159 ö	260
	TUHO2	259	431 ö	690
	TUHO3	180	271 ö	451
	TOTAL	540	861 ö	1401

(5)

(5) Ohjelma tulostaa havaintoaineiston hieman eri muotoisena taulukkona, kuin esimerkin esittelyssä on tehty.

/ FIT ASSOCIATION=3.

Tässä kapppaleessa saadaan yleiskäsitys taulukkoa kuvaavan mallin rakenteesta ja viitteitä siitä, mitkä täysiasteisen (saturated) mallin F=m+T+U+L+TU+LU+LT+LTU tekijöistä tarvitaan lopulliseen malliin.

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

Ts (9) Ts (8)

K-FACTOR	D.F.	LR CHISQ	PROB.	PEARSON CHISQ	PROB.
0-MEAN	11	803.11	0.00000	821.	0.00000
1	7	648.79	0.00000	672.	0.00000
2	2	0.96	0.61969	0.	0.61810
3	0	0.	1.	0.	1.

(6)

(6) Testi sille, ovatko kaikki astetta k korkeampaa astetta olevat tekijät nollia (ei-merkitseviä). Jos yhdysvaikutus ei ole merkitsevä, siinä olevat muuttujat eivät riipu toisistaan. Esimerkiksi toista astetta (kahden muuttujan yhdysvaikutustekijää) ylempää astetta olevia yhdysvaikutuksia ei mallissa tarvita, tässä tapauksessa tekijää LTU.

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

Ts (11)

K-FACTOR	D.F. (7)	LR CHISQ	PROB.	PEARSON CHISQ	PROB.
1	4	154.32	0.00000	149 10	0.00000
2	5	647.84	0.00000	671.44	0.00000
3	2	0.96	0.61969	0 96	0.61810

(7) Testi sille, ovatko kaikki astetta k olevat tekijät nollia. Esimerkiksi kaikki toisen asteen yhdysvaikutukset eivät ole nollia.

***** ASSOCIATION OPTION SELECTED FOR ALL TERMS OF ORDER LESS THAN R EQUAL TO 3

Ts (11) Ts (11)

EFFECT		PARTIAL ASSOCIATION D.F.	CHISQUARE	PROB	ITER	MARGINAL ASSOCIATION D.F.	CHISQUARE	PROB (9)	ITER
L.		1	5.71	0.0169					
T.		2	126.68	0.0000					
U.		1	21.93	0.0000					
LT.	(8)	2	0.13	0.9355	2	2	37.19	0.0000	2
LU.		1	203.56	0.0000	2	1	240.62	0.0000	2
TU.		2	370.02	0.0000	2	2	407.08	0.0000	2
LTU.		2	0.96	0.6197					

(8) Testi sille, onko tietty tekijä nolla sen mallin suhteen, jossa on mukana kaikki ko. astetta ja alempaa astetta olevat tekijät. Esimerkiksi yhdysvaikutus LT ei ole merkitsevä mallissa F=m+T+U+L+TU+LU+LT.

(9) Testi sille, onko tietty yhdysvaikutus merkitsevä itseensä sisältyvien muuttujien määrittelemässä marginaalitaulukossa. Esimerkiksi yhdysvaikutus LT on merkitsevä mallissa F=m+L+T+LT, joka kuvaa 2-suuntaista taulukkoa (14), joka on saatu summaamalla frekvenssit yli U-muuttujan luokkien.

Johtopäätökset:

Lopulliseen malliin otetaan mukaan muuttujat T,U ja L ja ehkä yhdysvaikutukset TU ja LU. Yhdysvaikutus LT on (9):n mukaan merkitsevä LxT-marginaalitaulukossa, mutta koska LT ei ole (8):n mukaan merkitsevä, niin yhdysvaikutusta LT ei luultavasti ole aihetta ottaa malliin.

/ FIT ALL.

2- ja 3-suuntaisissa taulukoissa voidaan ALL-parametrilla haluttaessa tulostaa kaikkien mahdollisten taulukkoa kuvaavien mallien yhteensopivuustestit (goodness of fit). Muutoin testattavat mallit on määriteltävä erikseen MODEL-parametrilla. Esimerkiksi /FIT MODEL=LU,TU.MODEL=TU,LT. tulostaisi toiseksi ja kolmanneksi alimmat rivit alla olevasta taulukosta.

MODEL	DF	Ts (9) LIKELIHOOD-RATIO CHISQ	PROB.	Ts (8) PEARSON CHISQ	PROB.	ITERATIONS
L.	10	797.40	0.0000	819.07	0.0000	1
T.	9	676.43	0.0000	688.42	0.0000	1
U.	10	781.18	0.0000	822.56	0.0000	1
L,T.	8	670.72	0.0000	687.85	0.0000	1
T,U.	8	654.50	0.0000	680.74	0.0000	1
U,L.	9	775.47	0.0000	811.91	0.0000	1
L,T,U.	7	648.79	0.0000	672.41	0.0000	1
LT.	6	633.53	0.0000	587.59	0.0000	1
LU.	8	534.85	0.0000	518.38	0.0000	1
TU.	6	247.42	0.0000	242.38	0.0000	1
L,TU.	5	241.71	0.0000	237.20	0.0000	1
T,LU.	6	408.17	0.0000	397.86	0.0000	1
U,LT.	5	611.60	0.0000	570.56	0.0000	1
LT,LU.	4	370.98	0.0000	367.72	0.0000	1
LU,TU. (10)	4	1.09	0.8958	1.09	0.8952	1
TU,LT.	3	204.52	0.0000	203.04	0.0000	1
LT,LU,TU.	2	0.96	0.6197	0.96	0.6181	6

(10) Malli F=m+T+U+L+TU+LU on suppein taulukkoa merkitsevästi kuvaava malli. Testeistä voidaan päätellä, että yhdysvaikutus LT ei ole merkitsevä. Testejä on tulkittava samaan tapaan kuin tavallisessa regressioanalyysissä: malliin ei ole mielekästä ottaa mukaan tekijöitä, jotka eivät ole merkitseviä, siitä huolimatta, että mallin selitysaste silloin kasvaa (tällöinhän selityksen "lisääntyminen" on täysin teknistä laatua). Vaikka esimerkiksi malli LT,LU,TU on testin mukaan merkitsevämmin sopusoinnussa havaintojen kanssa kuin malli LU,TU, niin se on näennäistä, koska LT ei merkitsevästi lisää selitystä. Tämä voidaan laskea yllä olevista testeistä muodostamalla tekijän LT merkitsevyyttä kuvaava

χ^2-testisuure (11):

$G_2^2 - G_1^2 = 1.09-0.96 = 0.13$, D.F.= 4-2 = 2,

mikä ei ole merkitsevä. Vertaa myös kohta (12).

/FIT MODEL=LTU.DELETE=SIMPLE.STEP=5.

Taulukon loglineaarinen malli voidaan määrätä myös askeltaen, lisäten tai poistaen tekijöitä MODEL-parametrilla määritellystä mallista, kunnes on päädytty malliin, jonka kaikki tekijät ovat merkitseviä. Tässä esimerkissä on lähdetty poistamaan tekijöitä täysiasteisesta mallista F=m+T+U+L+TU+LU+LT+LTU.

```
MODELS                                                    Ts (9)+
MODELS FORMED BY DELETING TERMS FROM MODEL —              Ts (11)
   LTU.
                                          LIKELIHOOD-RATIO          PEARSON
   MODEL                          D.F.    CHI-SQUARE    PROB        CHI-SQUARE    PROB
   TU,LU,LT.                      2         0.96        0.6197        0.96        0.6181
     DIFF. DUE TO DELETING  LTU.  2         0.96        0.6197
STEP  1.     BEST MODEL FOUND IS —
        TU,LU,LT.
```

Ts (9)
Ts (10)

(11)

(11) Askeleessa 1 on päädytty malliin F=m+T+U+L+TU+LU+LT.

```
   TU,LU.                         4         1.09        0.8958        1.09        0.8952
     DIFF. DUE TO DELETING  LT.   2         0.13        0.935
   TU,LT.                         3       204.52        0.0000      203.04        0.0000
     DIFF. DUE TO DELETING  LU.   1       203.56        0.0000
   LU,LT.                         4       370.98        0.0000      367.72        0.0000
     DIFF. DUE TO DELETING  TU.   2       370.02        0.000
STEP  2.     BEST MODEL FOUND IS —
       TU,LU.
```

Ts (9)
Ts (11)

(12)

(12) Testisuure (9) osoittaa, että malli TU,LU on merkitsevä. Testisuure (11) = 1.09-0.96=0.13, D.F.=4-2=2 osoittaa, että yhdysvaikutus LT ei ole merkitsevä mallissa TU,LU,LT,vrt. (10) Askeleessa 2 päädytään malliin TU,LU: F=m+T+U+L+TU+LU.

```
   TU,L.                          5       241.71        0.0000      237.20        0.0000
     DIFF. DUE TO DELETING  LU.   1       240.62        0.0000
   T,LU.                          6       408.17        0.0000      397.86        0.0000
     DIFF. DUE TO DELETING  TU.   2       407.08        0.0000
STEP  3.     BEST MODEL FOUND IS —
       T,LU.

 STEPPING STOPS DUE TO CRITERION PROBABILITY (     0.050).
```

(13)

(13) Askeleen 3 perusteella mallista TU,LU ei enää voida poistaa yhtään tekijää. Askeleessa saatu paras malli T,LU: F=m+T+U+L+LU ei enää ole merkitsevä, joten edellisessä askeleessa (STEP 2:ssa) saatu malli kuvaa aineistoa parhaiten.

Askeltavan analyysin lopputulos on sama kuin edellisessä /FIT-kappaleessa. Tulokset vahvistavat toisiaan.

/PRINT EXPECTED.DIFFERENCES.LAMBDA.

Vertaamalla mallilla laskettuja (expected) frekvenssejä havaittuihin frekvensseihin paljastuu joskus, missä kohdin malli ei vastaa havaintoja.

Ts (9) LIKELIHOOD-RATIO — Ts (8) PEARSON

MODEL	D.F.	CHI-SQUARE	PROB	CHI-SQUARE	PROB
LU,TU.	4	1.09	0.8958	1.09	0.8952

***** EXPECTED VALUES USING ABOVE MODEL

UUDTAPA	TUHO	LAJI TAIMLAJ1	TAIMLAJ2	TOTAL
UUDTAPA1	TUHO1	447.7	201.3	649.0
	TUHO2	249.0	112.0	361.0
	TUHO3	106.2	47.8	154.0
	TOTAL	803.0	361.0	1164.0
UUDTAPA2	TUHO1	100.2	159.8	260.0
	TUHO2	266.0	424.0	690.0
	TUHO3	173.8	277.2	451.0
	TOTAL	540.0	861.0	1401.0

(14)

(14) Mallilla LU,TU: F=M+T+U+L+TU+LU lasketut frekvenssit.

***** DIFFERENCES BETWEEN OBSERVED AND EXPECTED USING ABOVE MODEL

UUDTAPA	TUHO	LAJI TAIMLAJ1	TAIMLAJ2
UUDTAPA1	TUHO1	0.3	-0.3
	TUHO2	3.0	-3.0
	TUHO3	-3.2	3.2
UUDTAPA2	TUHO1	0.8	-0.8
	TUHO2	-7.0	7.0
	TUHO3	6.2	-6.2

(15)

(15) Taulukossa on havaittujen (5) ja mallilla laskettujen (14) frekvenssien erotukset. Vastaavuus on hyvä.

```
THE ABOVE MODEL IS DIRECT

ESTIMATES OF THE LOG-LINEAR PARAMETERS (LAMBDA) IN THE MODEL ABOVE
THETA(MEAN)  5.1895  (16)

*****  ESTIMATES OF THE LOG-LINEAR PARAMETERS (LAMBDA) IN THE MODEL

              TUHO
             ------
          TUHO1    TUHO2    TUHO3
        ---------------------------
         0.083    0.278   -0.361   (17)

*****  (16)  RATIO OF THE LOG-LINEAR PARAMETER ESTIMATE TO ITS STANDARD ERROR

                    TUHO
                   ------
                TUHO1    TUHO2    TUHO3
              ---------------------------
                2.683    9.469  -10.268  (17)

*****  ESTIMATES OF THE LOG-LINEAR PARAMETERS (LAMBDA) IN THE MODEL

UUDTAPA       LAJI
-------      ------
          TAIMLAJ1 TAIMLAJ2
        ------------------
UUDTAPA1   0.317   -0.317
UUDTAPA2  -0.317    0.317

*****  (16)  RATIO OF THE LOG-LINEAR PARAMETER ESTIMATE TO ITS STANDARD ERROR

          UUDTAPA       LAJI
          -------      ------
                   TAIMLAJ1 TAIMLAJ2
                 ------------------
          UUDTAPA1  15.101  -15.101
          UUDTAPA2 -15.101   15.101
```

(16) Frekvenssien ohella myös mallin parametrien arvoja vertailemalla voidaan arvioida vaikutusten voimakkuutta ja rakennetta.
Mallin $F_{ijk}=m+T_i+U_j+L_k+(TU)_{ij}+(LU)_{jk}$ parametrien arvoja ovat esimerkiksi

$$m = 5.1895$$

$$T_1 = 0.083$$

$$(UL)_{21} = -0.317$$

(17) Ohjelma tulostaa myös likimääräisen testisuureen (parametrin arvo/parametrin hajonta) parametrien merkitsevyyden testaamista varten. Parametrin merkitsevyystestissä peukalosääntönä on: jos testisuureen itseisarvo on ≈ 2 tai suurempi, parametri poikkeaa merkitsevästi nollasta. Esimerkiksi vaikutus $T_3=-0.361$ poikkeaa varmasti nollasta, koska $|-10.268|>>2$.

Kirjallisuutta:

Anderson, S. & Auquier, A. & Hauck, W.W. & Oakes, D. & Vandaele, W. & Weisberg, H.I. 1980. Statistical methods for comparative studies. Wiley, New York. 289 p.

BMDP Statistical software manual 1985 printing. 1985. (Dixon, W. J. ed.). University of California Press, Berkeley. 733 p.

Cox, D. R. 1970. Analysis of binary data. Chapman and Hall, London. 142 p.

Ekholm, A. 1983. Frekvenssiaineiston analyysi. Sosiaalilääketieteellinen Aikakauslehti 1983:20:89-96.

Lindgren, B. W. 1976. Statistical theory. Macmillan, New York. 614 p.

METSÄNTUTKIMUSLAITOS

Matemaattinen osasto
Osoite: PL 37, 00381 HELSINKI (Kornetintie 8) ja
Unioninkatu 40 A, 00170 HELSINKI
Puhelin: (90) 556 276 ja
(90) 661 401

Hari, Pertti, vs. professori

Klippi, Lea, tutkimussihteeri

Menetelmät
Häkkinen, Risto, matemaatikko
Heinonen, Jaakko, tutkija (Joensuun tutkimusasema)
Sievänen, Risto, tutkija

Atk
Pöntinen, Jukka, atk-päällikkö
Herrala-Ylinen, Helena, tutkija
Kaila, Erkki, tutkija (Rovaniemen tutkimusasema)
Kinnunen, Hilkka, tutkija (Rovaniemen tutkimusasema)
Mäkinen, Markku, tutkija
Salmi, Veli-Pekka, atk-suunnittelija
Snellman, Carl-Gustaf, tutkija
Granlund, Hilkka, pääoperaattori
Palviainen, Pertti, tutkimusapulainen (Rovaniemen tutkimusasema)
Soimula, Maire, operaattori

Metsätilasto
Uusitalo, Matti, tutkija
Aarne, Martti, tutkija
Lehto, Kari, tutkija
Valli, Pasi, tutkija
Leppäkumpu, Tuula, toimistosihteeri
Kämäräinen, Paula, toimistosihteeri

Metsäverotus
Rauskala, Raimo, vanhempi tutkija
Kakkuri, Eero, tutkija
Kulju, Irma, toimistosihteeri
Mäkinen, Kaija, ohjelmoija

Sivulliset tutkijat
Kallio, Markku, professori
Rytkönen, Antti, metsänhoitaja

Matemaattisella osastolla ilmestyneet Metsäntutkimuslaitoksen tiedonantoja -sarjan viimeisimmät julkaisut:

nro 149 Pertti Hari, Kullervo Kuusela, Pentti K. Räsänen, Risto Seppälä. Metsäntutkimukseen liittyvistä kehityssuunnista. 38 s. 1984.

nro 152 Eero Kakkuri. Yksityismetsänomistajien puun kasvatuksen kulut vuosina 1981 ja 1982. 17 s. 1984.

nro 157 Erkki Kaila ja Markku Taipale. Tutka-tiedonhallintaohjelmisto Tietokannan muodostus ja käyttö. 113 s. 1984.

nro 176 Raimo Rauskala. Forest taxation and roundwood supply in Finland. 12 s. 1985.

nro 183 Staffan Ringbom. Virkesproduktionens totala lönsamhet och dess mätning. 32 s. 1985.

nro 191 Raimo Rauskala. Kunnittaiset kantohinnat ja puukuutiometrin bruttoarvot hakkuuvuonna 1983/84. 44 s. 1985.

nro 194 Heinonen, J., Penttinen, A., Salminen, S., Tomppo, E. Spatiaalisen tilastotieteen soveltaminen metsäntutkimukseen. 129 s. 1985.

nro 223 Raimo Rauskala. Kunnittaiset kantohinnat ja puukuutiometrin bruttoarvot hakkuuvuonna 1984/85. 55 s. 1986.

nro 224 Eero Kakkuri. Puun hintojen vaihtelu kuntien sisällä hakkuuvuonna 1980/81. 22 s. 1986.

nro 240 Eero Kakkuri. Yksityismetsänomistajien puun kasvatuksen kulut vuosina 1983 ja 1984. 22 s. 1986.

nro 251 Eija Virtanen (toim.). BIB-viitetietokantaohjelmisto. Version 2.1 käyttöohje. 62 s. 1987.

nro 252 Hilkka Kinnunen. Metlan sarjat viitetietokantana. 24 s. 1987.

nro 254 Kari Lehto. Turvallisuusnäkökohdista ja suojauksista Metsäntutkimuslaitoksen atk-järjestelmissä. 55 s. 1987.

nro 265 Raimo Rauskala. Kunnittaiset kantohinnat ja puukuutiometrin bruttoarvot hakkuuvuonna 1985/86. 59 s. 1987.

nro 269 Kimmo Linnilä. Tilastollinen tietojenkäsittely mikrotietokoneilla. 29 s. 1987.

ISBN 951-40-0831-6
ISSN 0358-4283