ASSOCIATION MODELING FOR SPARSE PRESENCE-ABSENCE DATA

Alasford M. Ngwengwe

and

Charles E. McCulloch

Biometrics Unit, Cornell University, Ithaca, NY

BU-1061-M

November, 1989

ASSOCIATION MODELING FOR SPARSE PRESENCE-ABSENCE DATA

Alasford M. Ngwengwe

and

Charles E. McCulloch

Biometrics Unit, Cornell University, Ithaca, New York 14853, U.S.A.

Key Words: association structure; Bernoulli random variable; conditional distribution; correlation coefficient; hypergeometric distribution; joint density; log-linear model; multivariate binary distribution;

SUMMARY

We model the association structure of multivariate binary variables using a model which naturally incorporates the parameter restrictions of the standard log-linear model. We propose some techniques for testing association for sparse data, through a conditioning argument. One of the test statistics is equivalent to the Mantel-Haenzsel statistics. We evaluate their performance through simulation and illustrate their use on data from the Forty-seventh American Breeding Birds Census.

1. Introduction

There are many situations in scientific investigation where a vector $\underline{Y}$ of responses is vector of binary random variables, i.e., $Y_j = 0$, or 1, for $j = 1, 2, \ldots, M$ is obtained on each sampled individual, experimental unit, or sampling unit. Examples include ecology where Y_j represents the presence-absence of the j^{th} species, in medical diagnosis where the data may take the form of presence versus the absence of symptoms. Other examples occur in taxonomy where the presence or absence of each of M characteristics is noted, in the evaluation of the performance of the equipment where each of the M components performs or fails, in educational testing where each of the M questions generates a yes-no response.

In many of the cases, especially if M is large, the data will be quite sparse and will not support the usual asymptotic inference for the log-linear models. In such cases an appropriate starting point may be to consider models for pairwise association which are flexible with sparse data. We propose several such techniques, evaluate their performance with sparse data sets and illustrate their use on a real data set. We consider modeling the association structure by directly modeling the multivariate binary distribution. We look at the genuinely multivariate situation in which there are several binary variables and wish to model the association between these variables and not just the dependence of one variate on others, as is the case in the classical logistic model. Cox (1972) briefly reviewed eight kinds of models. One of them is the multivariate logistic model which we will refer to as

the Cox model. We describe the Cox model in the next section.

2. Formulation of Cox model

The Cox model is a log-linear model for binary data which naturally incorporates the constraints of the parameters. For example let Y_1, Y_2 and Y_3 be Bernoulli random variables with joint probability function P_{ijk}, the probability that $Y_1 = i$, $Y_2 = j$, and $Y_3 = k$, i; j; k = 0 or 1 and the sum of the P_{ijk}s is unity. The standard log-linear model for the probabilities can be represented by a single equation, i.e.,

$$\log\left[P_{ijk}\right] = U + (-1)^{i+1}U_1 + (-1)^{j+1}U_2 + (-1)^{k+1}U_3 +$$

$$(-1)^{i+j}U_{12} + (-1)^{i+k}U_{13} + (-1)^{i+j}U_{23} + (-1)^{i+j+k+1}U_{123}. \qquad (2.1)$$

The above equation (2.1) can be written in a compact form by making the following transformation, $Z_j = 1 - 2Y_j$, j =1, 2, Z_j takes values 1 and -1, and writing α_j for U_j, γ_{jk} for U_{jk}, γ_{123} for U_{123}, and $-\varphi(\alpha,\gamma)$ for U we get the representation in Cox (1972), which will be adopted and is given by

$$\log\left\{P\left[Z_1=z_1, Z_2=z_2, Z_3=z_3\right]\right\} = \alpha_1 z_1 + \alpha_2 z_2 + \alpha_3 z_3 + \gamma_{12} z_1 z_2 +$$

$$\gamma_{13} z_1 z_3 + \gamma_{23} z_2 z_3 + \gamma_{123} z_1 z_2 z_3 - \varphi(\alpha,\gamma). \qquad (2.2)$$

For other representations see Nerlove and Press (1973), Rosner (1984), and Bonney (1987). Let $\underline{Y}$ be a vector of Bernoulli random variables and transform the Y_j's to Z_j's as given above and let Ω

be the set of all 2^M possible values of $\underline{Z}$. In the Cox model the log of the probability of $\underline{Z} = \underline{z}$ is

$$\log\left\{P\left[\underline{Z} = \underline{z}\right]\right\} = \sum_{j=1}^{M} \alpha_j z_j + \sum_{j<k}^{M}\sum^{M} \gamma_{jk} z_j z_k + \sum_{j<k<t}^{M}\sum^{M}\sum^{M} \gamma_{jkt} z_j z_k z_t$$

$$+ \ldots \gamma_{12\ldots M} z_1 z_2 \cdots z_M - \varphi(\alpha,\gamma) \qquad (2.3)$$

where $\varphi(\alpha,\gamma)$ is the normalizing constant so that the sum of the probabilities over the set Ω is one. If we assume that all the second and higher order interaction terms in (2.3) are zero, the joint density of the Z_j's is given by

$$P\left[\underline{Z} = \underline{z}\right] = \exp\left[\sum_{j=1}^{M} \alpha_j z_j + \sum_{j<k}^{M}\sum^{M} \gamma_{jk} z_j z_k - \varphi(\alpha,\gamma)\right] \qquad (2.4)$$

where

$$\exp\left[\varphi(\alpha,\gamma)\right] = \sum_{\underline{z} \in \Omega} \exp\left[\sum_{j=1}^{M} \alpha_j z_j + \sum_{j<k}^{M}\sum^{M} \gamma_{jk} z_j z_k\right]$$

this is the model we will assume for the rest of the discussion. This will be an appropriate model for situations where pairwise association predominates. The α_j's are referred to as the main effects and the γ_{jk}'s as the interaction terms. α_j (j = 1, 2, . . ., M) in the model is one-half the average of the log odds for the j^{th} species over the presence-absence combination of the other M-1 species, while γ_{jk} is one-fourth the log odds ratio for the j^{th} (k^{th}) species to the k^{th} (j^{th}) species. Nerlove and Press (1973) and Bonney (1987) give expressions of the parameters in terms of the log of the cell probabilities for saturated models. The mean and covariance structure both depend on $\underline{\alpha}$ and $\underline{\gamma}$. For example for M

= 3 with $\alpha_j = \alpha$, and $\gamma_{jk} = \gamma$, the common correlation coefficient is given by

$$\text{corr}\left[Z_j, Z_k\right] = \frac{\left[\theta^2 + \theta K_1 - K_1 - 1\right]}{\left[\theta^2 + (2K_1 + K_2)\theta + 2K_1 + 5\right]} \qquad (2.5)$$

where $\theta = \exp(4\gamma)$, $K_1 = \exp(2\alpha) + \exp(-2\alpha)$ and, $K_2 = \exp(4\alpha) + \exp(-4\alpha)$. The derivative of (2.5) with respect to θ is positive so that the correlation coefficient is a strictly monotone increasing function of θ (or γ), see Figure I which is graph of (2.5).

There are many special cases of this multivariate distribution which are of interest for modeling the association structure in multivariate binary variables. For example, when all the γ_{jk}'s are all zero, we have independence. Another special case of the Cox model which could be of interest is the analogue of equal correlation models in normal theory suggested by Cox (1972), where one considers the case of all γ_{jk}s equal and all higher order interactions are zero. Note that there is equal correlation if all the α_j's are equal and all the γ_{jk}'s are also equal.

For all γ_{jk}'s equal to zero we have independence and the common correlation coefficient is a monotone increasing function in γ when all the γ_{jk}'s are equal to γ with the α_j's also equal as Figure I shows for M = 3. Therefore assuming a common gamma, its estimate or function of it is a reasonable measure of association. The α_j's can be treated as nuisance parameters as we are more interested in the correlation or interaction of the variables.

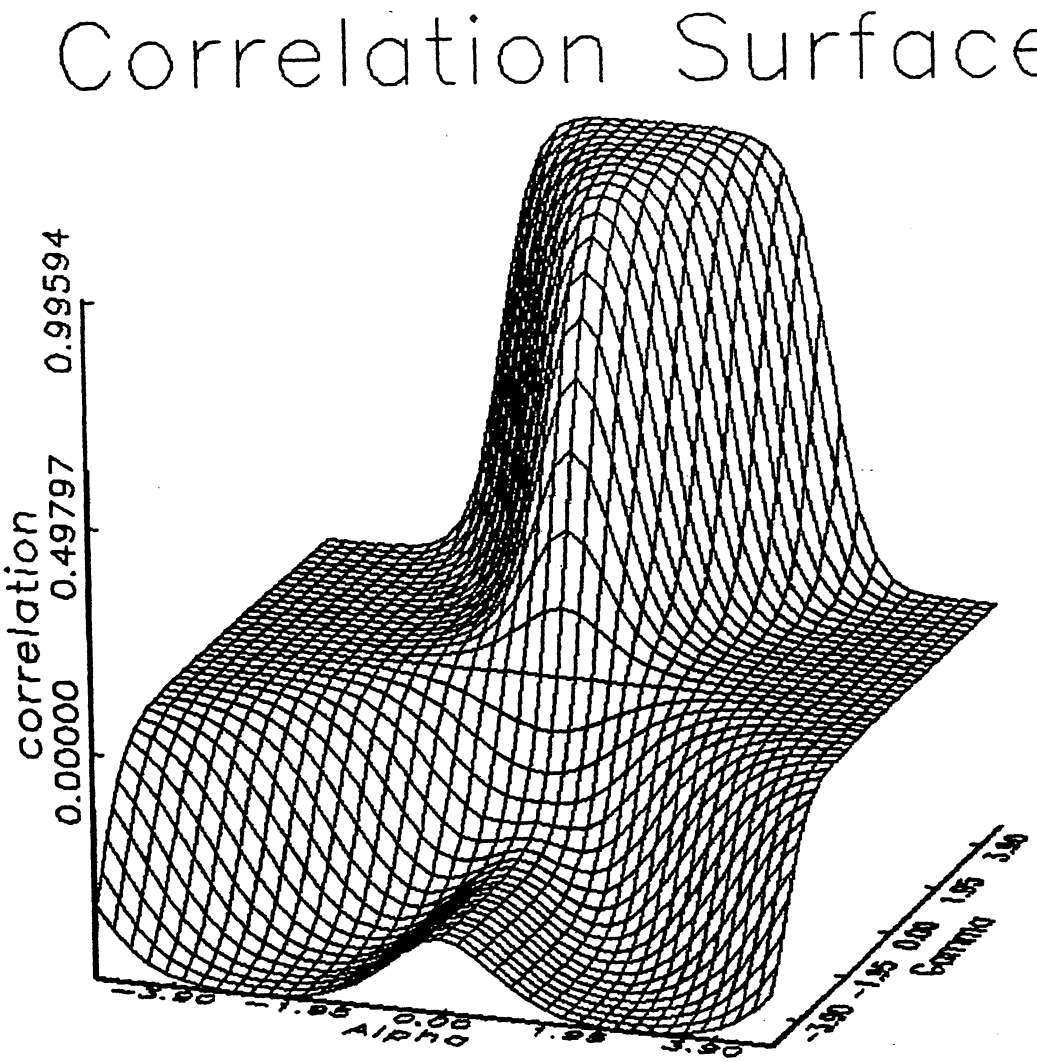

Figure I : Correlation surface with alphas all equal for M = 3, and gammas all equal, i.e. $\gamma_{jk} = \gamma$ with $j < k$, equation (2.5).

In section 3 we eliminate these nuisance parameters through a conditioning argument.

3. Exact and approximate inference

3.1 Exact inference

The sufficient statistic of $(\underline{\alpha},\underline{\gamma})$ for a random sample of size N from the Cox model (2.4) is $(\underline{T},\underline{S})$ where

$$T_j = \sum_{i=1}^{N} Z_{ij} \quad \text{and} \quad S_{jk} = \sum_{i=1}^{N} Z_{ij}Z_{ik} .$$

Following Cox (1970), Tritcher (1984), Hirji, Mehta, and Patel (1988), and Hirji, Mehta, and Tsiatis (1989) to name a few, the joint density for the sufficient statistics $(S,\underline{T})$ is given by

$$P\left[\underline{S} = \underline{s},\underline{T} = \underline{t}\right] = C_N\left[\underline{s},\underline{t}\right]\exp\left[\underline{\alpha}^t\underline{t} + \underline{\gamma}^t\underline{s} - N\varphi(\alpha,\gamma)\right]$$

where $C_N(\underline{s},\underline{t})$ is the number of distinct $\underline{Z}$ vectors yielding the vector $(\underline{s},\underline{t})$ for the sufficient statistics. Transforming back to the zero-one binary case, i.e.

$$S_{jk} = 4U_{jk} - 2N_j - 2N_k + N \quad \text{and} \quad T_j = 2N_j - N, \tag{3.1}$$

where U_{jk} is the number of trials resulting in successes for both treatments in N independent trials, N_j is the number of successes for the j^{th} treatment and N_k is for the k^{th} treatment. Then we have $P(\underline{S} = \underline{s},\underline{T} = \underline{t}) = P(\underline{U} = \underline{u},\underline{N} = \underline{n})$ and $C_N(\underline{s},\underline{t}) = C_N(\underline{u},\underline{n})$. It is easy to show that conditional distribution of $\underline{U}$ given $\underline{N} = \underline{n}$ is

$$P\left[\underline{U}=\underline{u}\,\middle|\,\underline{N}=\underline{n}\right] = \frac{C_N\left[\underline{u},\underline{n}\right]\prod_{j<k}^{M}\left[\Psi\left[\gamma_{jk}\right]\right]^{u_{jk}}}{\sum_{\underline{w}} C_N\left[\underline{w},\underline{n}\right]\prod_{j<k}^{M}\left[\Psi\left[\gamma_{jk}\right]\right]^{w_{jk}}} \qquad (3.2)$$

where $\underline{w}$ is an index vector ranging over the values taken by $\underline{U}$, and $\Psi(\gamma_{jk}) = \exp(4\gamma_{jk})$.

For M = 2, (3.2) reduces to the distribution of Fisher's exact test statistic for testing independence in a 2 x 2 table, a non-central hypergeometric with odds-ratio parameter $\Psi(\gamma)=\exp(4\gamma)$ as the non-centrality parameter. However, with the restriction that the marginal probabilities are equal, equation (3.2) is density for a non-central hypergeometric distribution with non-centrality parameter $\Psi(\gamma)/4$, and under the null hypothesis it is still non-central hypergeometric, but the non-centrality is a known value, 1/4.

3.2 Approximate inference using Cox model

From equation (2.4) it can be shown that under $H_0 : \underline{\gamma} = 0$ for every i, the Z_{ij}'s are also independent, it follows that the Y_{ij}'s are independent. From the above results we get the conditional means, variances and covariances of the U_{jk}'s under H_0 are respectively

$$E_{H_0}\left[U_{jk}\,\middle|\,\underline{N}=\underline{n}\right] = \frac{n_j n_k}{N}\,, \qquad (3.3)$$

$$\mathrm{Var}_{H_0}\left[U_{jk}\,\middle|\,\underline{N}=\underline{n}\right] = \frac{n_j\left[N-n_j\right]n_k\left[N-n_k\right]}{N^2(N-1)} \tag{3.4}$$

$$\underset{k\neq v}{\mathrm{Cov}_{H_0}}\left[U_{jk},U_{jv}\,\middle|\,\underline{N}=\underline{n}\right] = \frac{n_k n_v}{N^2}\left\{n_j + n_j\left[n_j-1\right] - n_j^2\right\} = 0,$$

and similarly the conditional covariance of U_{jk} and U_{vt} is zero for $j \neq v$ and $k \neq t$, so under H_0, the U_{jk}'s conditional on the N_j's are uncorrelated and U_{jk} has mean and variance of a hypergeometric variate with parameters N, n_j and n_k.

We construct a test statistic Q, based on Mahalanobis distance of the statistic $\underline{U}$ from its conditional mean under the null hypothesis which upon simplification is given by

$$Q = (N-1)\sum_{j<k}^{M}\sum^{M}\frac{\left[NU_{jk} - n_j n_k\right]^2}{n_j\left[N-n_j\right]n_k\left[N-n_k\right]}\,. \tag{3.5}$$

Q is a sum of squares M(M-1)/2 uncorrelated standardized random variables (under H_0). Therefore, under H_0, Q is distributed approximately as chi-square with M(M-1)/2 degrees of freedom. The statistic Q is (N-1)/N times the sum of the usual chi-square statistics for testing independence in M(M-1)/2 independent 2 x 2 tables. An alternative test statistic is the normal approximation, the sum of the variables in Q without squaring them. This test statistic is given by

$$Q^{*} = \left[\frac{2(N-1)}{M(M-1)}\right]^{1/2} \sum_{j<k}^{M}\sum^{M} \frac{\left[NU_{jk} - n_j n_k\right]}{\left[n_j\left[N-n_j\right]n_k\left[N-n_k\right]\right]^{1/2}} . \qquad (3.6)$$

Under H_0, Q^{*} has an approximate standard normal distribution.

If we start with the restriction that γ_{jk}s are equal to a common value, γ, then the sufficient statistics for $(\gamma, \underline{\alpha})$ is $(U_{com}, \underline{N})$ where U_{com} is the sum of the U_{jk}'s. Therefore we can use Q_{com} defined below (based on Mahalanobis distance as Q) to test that the common γ is zero and is given by

$$Q_{com} = \frac{(N-1)\left\{\sum_{j<k}^{M}\sum^{M}\left[NU_{jk} - n_j n_k\right]\right\}^{2}}{\sum_{j<k}^{M}\sum^{M} n_j\left[N-n_j\right]n_k\left[N-n_k\right]} . \qquad (3.7)$$

Under H_0, Q_{com} is distributed approximately as chi-square with 1 degree of freedom. We note that Q_{com} is the Mantel-Haenszel statistic, Mantel and Haenszel (1959), and Mantel (1963), for testing that the common odds ratio is one, in M(M-1)/2 independent 2 x 2 tables. Note, however, that the data for our situation does not come as sets of 2 x 2 tables. Alternatively we could test the hypothesis that the common gamma is zero by the statistic Q^{*}_{com}, the signed square root of Q_{com}, which is approximately standard normal under H_0 (Q^{*}_{com} rejects if and only if Q_{com} does reject). The test Q could be used when the direction of independence is not important and test Q^{*} and Q_{com} (or Q^{*}_{com}) could be used when the

direction of independence (negative or positively correlated on the average) is important. If the deviation $NU_{jk} - n_j n_k$ was large positive and another large negative, then for test Q^* and Q_{com} they would cancel out, whereas for test Q they would add up.

The exact null distributions of Q, Q^* and Q_{com} were computed for selected $\underline{N}$ vectors for a gamma structure for which the exact distribution can be calculated with N = 21 for M = 3 and M = 4. The exact and approximate cumulative distribution functions (CDFs) were compared to check the accuracy of the approximations. The chi-square approximations for Q and Q_{com} and the normal approximation for Q^* are quite good. The approximations get better as the number of sample points of $\underline{U}$ increases (which is a function of the vector $\underline{N}$ and the sample size). Exact inference can be done for Q_{com}, using the results of Mehta, Patel and Gray (1985), and the program STATXACT (Gajjar, Hilton, Mehta, Patel, Senchaudhuri, and Walsh; 1989).

4. Results

4.1 Some simulation results

Some simulation studies were done to examine the performance of the test statistics Q, Q^*, and Q_{com} in terms of their power using the approximate null distributions for the cut-off points. For several gamma structures the power functions of Q, Q^*, and Q_{com} were examined for M = 3 and M = 4 with the alpha configuration.

$$\alpha_1 = -0.69 \ , \ \alpha_2 = -0.42 \ , \ \alpha_3 = -0.20 \quad \text{and} \quad \alpha_4 = 0 \qquad (4.1)$$

giving the marginal probabilities under independence as

$$P_{H_0}\left[Z_1 = 1\right] \doteq 0.2 \quad , \quad P_{H_0}\left[Z_2 = 1\right] \doteq 0.3 \quad ,$$

$$P_{H_0}\left[Z_3 = 1\right] \doteq 0.4 \quad \text{and} \quad P_{H_0}\left[Z_4 = 1\right] = 0.5 \ .$$

The alpha structure given in (4.1) was chosen with an ecological application in mind, in attempt to model the occurrence of species in the species-site problem.

For the following gamma structures, simulated power functions are graphed for sample size N = 21 and a thousand trials. All the programs used in this paper were written in GAUSS, a programming language, Edlefsen and Jones (1987).

$\gamma_{jk} = \gamma$, $j < k$, $j;k = 1, 2,$ and 3, (Figure II)

$\gamma_{12} = \gamma_{23} = \gamma_{34} = \gamma$ and $\gamma_{13} = \gamma_{14} = \gamma_{24} = -\gamma$, (Figure III)

$\gamma_{12} = 0.2 + \gamma$, $\gamma_{13} = 0.2 - \gamma$, and $\gamma_{23} = 0.2$, (Figure IV)

for γ values between -2 and 2. The hypothesis tested is H_0: $\underline{\gamma} = 0$ vs H_1: $\underline{\gamma} \neq 0$, at significance level 0.05. Recall that Q and Q^* test that all the gammas are equal to zero and Q_{com} tests that the common gamma is zero. From the simulated power functions (Figures II through IV) Q and Q_{com} do perform as expected, i.e., Q_{com} has better power than Q when all the gammas are equal, while Q has better power than Q_{com} when the gammas are not all equal. The performance of Q^* behaves the same as Q_{com} in terms of the power for the gamma structures investigated for M = 3 and M = 4, that is, Q^* is also testing that the common gamma is zero. Figure II

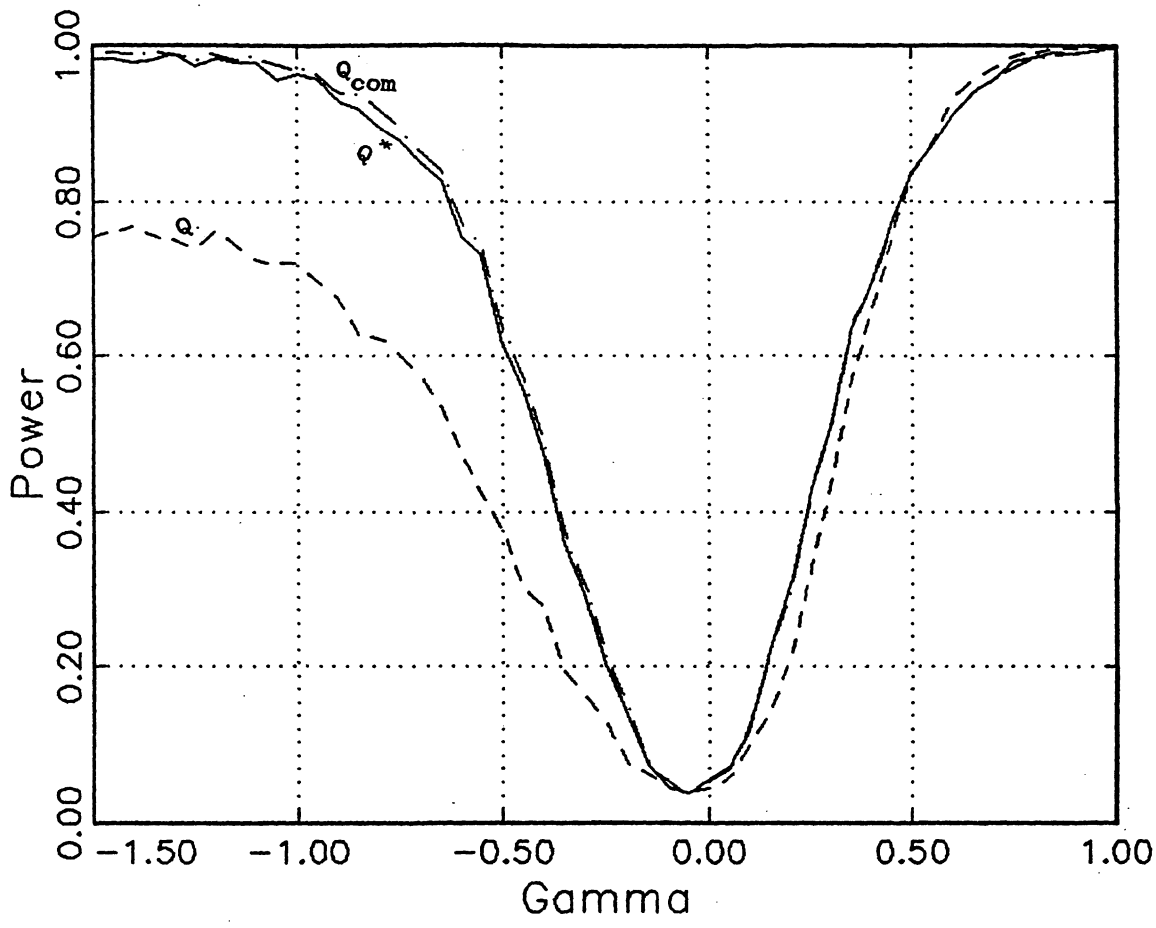

Figure II : Power functions of Q, Q^* and Q_{com}, with $\gamma_{jk} = \gamma$, for $j < k$; $j,k = 1, 2, 3$.

demonstrates the superior performance of Q^* and Q_{com} over Q for detecting non independence when the gammas are all equal, especially when the common gamma is negative. Figure III shows the superior power of Q over Q^* and Q_{com} in detecting non independence when the gammas are not all equal. Q is not uniformly more powerful than Q^* and Q_{com}, that is, one does not lose much power

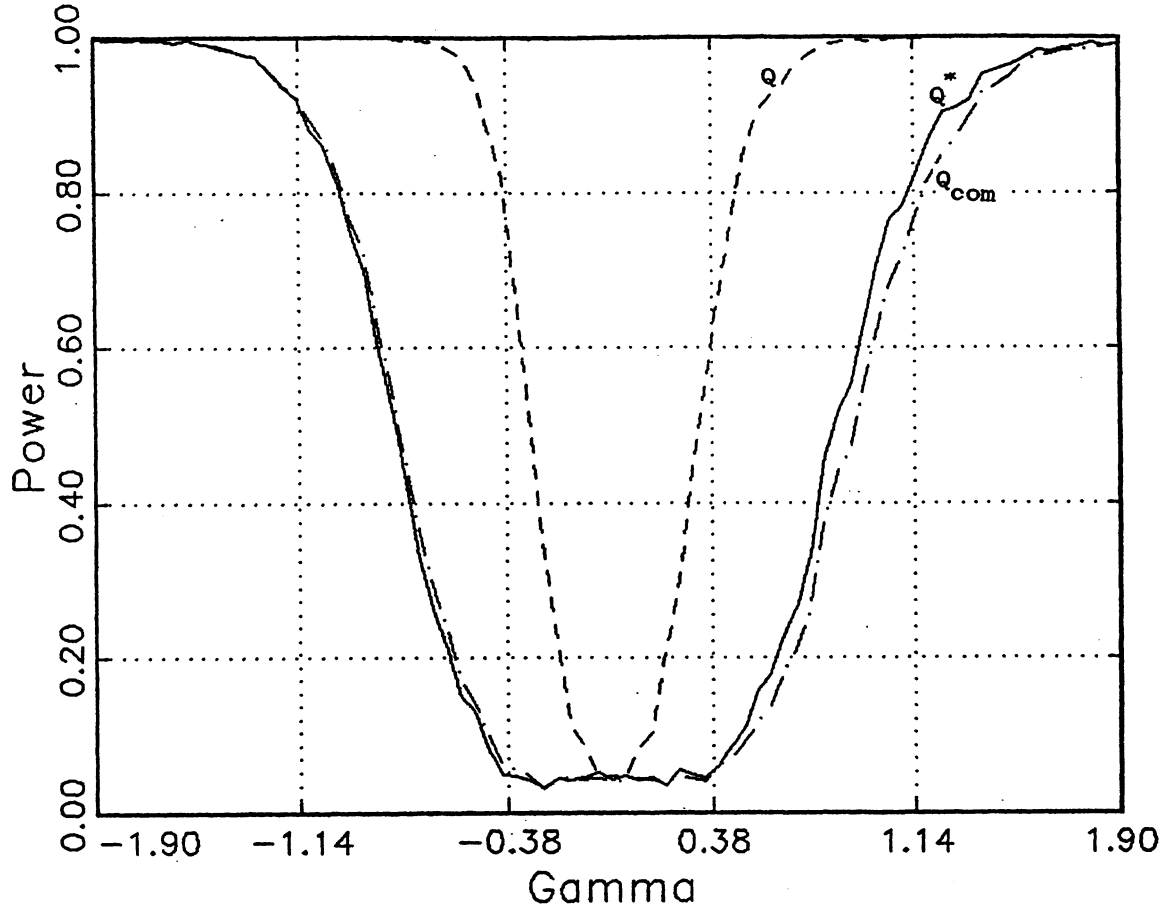

Figure III : Power functions of Q, Q^* and Q_{com}, with $\gamma_{12} = \gamma_{23} = \gamma_{34} = \gamma$ and $\gamma_{13} = \gamma_{14} = \gamma_{24} = -\gamma$.

by using Q^* or Q_{com} instead of Q when the gammas are not all equal for some gamma structures. Figure IV demonstrates the complete lack of power of Q^* and Q_{com} to detect non independence for certain gamma structures while the power of Q is relatively unchanged. Q^* and Q_{com} have very small power for all values of γ between -2 and 2.

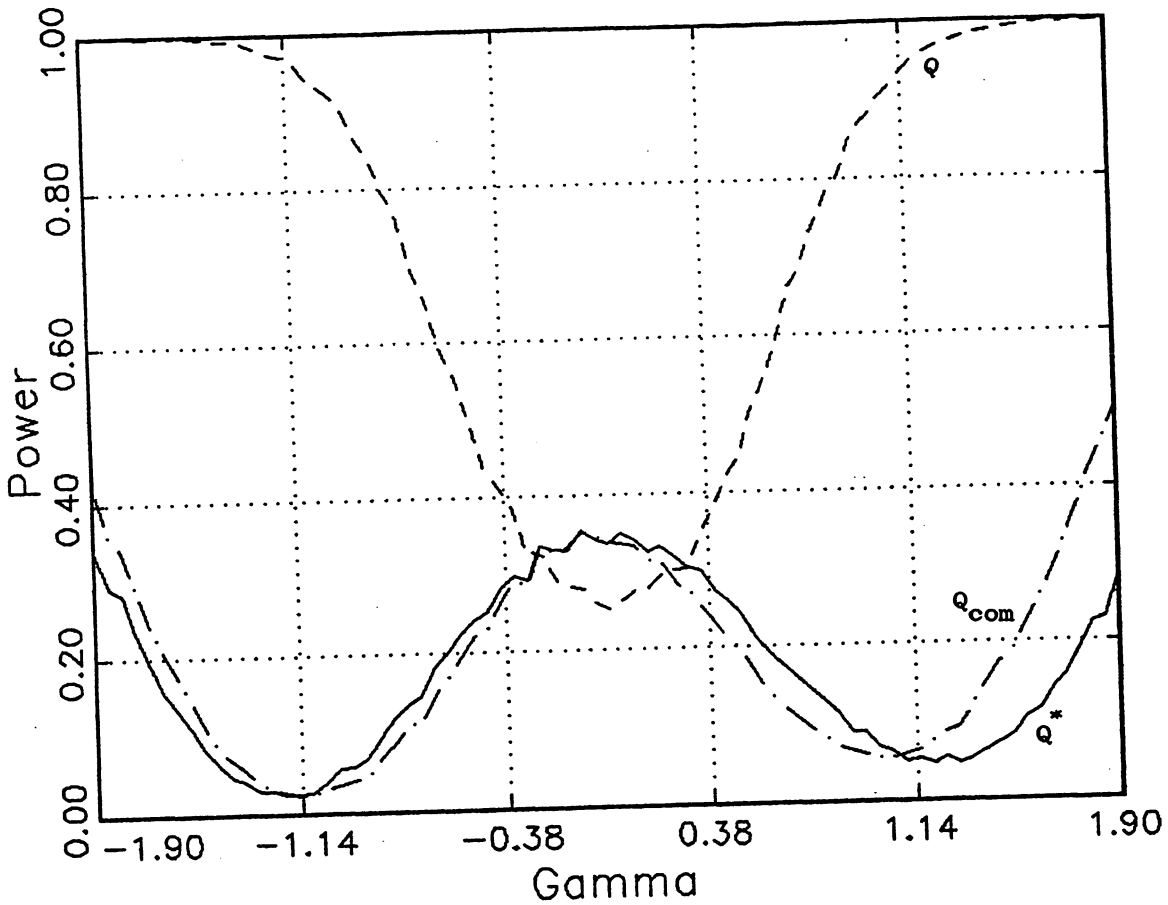

Figure IV : Power functions of Q, Q^* and Q_{com}, with $\gamma_{12} = 0.2 + \gamma$, $\gamma_{13} = 0.2 - \gamma$ and $\gamma_{23} = 0.2$.

While Q^* and Q_{com} have reasonable power for unequal gammas in a number of cases, their power is very small for gamma structures which induce negative and positive association of almost equal magnitude. It is easy to see what is happening by examining the 2 x 2 interaction tables in terms of the probabilities for the given gamma structure. With the configuration of $\underline{\alpha}$ given in equation

Table I : The 2 x 2 interaction tables along with the marginal probabilities for the gamma structure $\gamma_{12} = 0.2 + \gamma$, $\gamma_{13} = 0.2 - \gamma$, and $\gamma_{23} = 0.2$ used for figure IV.

γ	Z_1Z_2	Z_1Z_3	Z_2Z_3	$Z_1 = 1$	$Z_2 = 1$	$Z_3 = 1$
-2	0.01 0.26 0.66 0.07	0.26 0.01 0.01 0.72	0.01 0.66 0.26 0.07	0.268	0.668	0.268
-1.2	0.02 0.24 0.51 0.23	0.20 0.03 0.04 0.73	0.06 0.49 0.20 0.25	0.231	0.528	0.239
-0.4	0.06 0.14 0.27 0.53	0.14 0.09 0.12 0.65	0.11 0.23 0.15 0.51	0.171	0.299	0.231
0	0.08 0.11 0.17 0.64	0.09 0.10 0.22 0.59	0.12 0.14 0.19 0.55	0.151	0.217	0.282
0.4	0.09 0.08 0.10 0.73	0.06 0.11 0.35 0.48	0.11 0.09 0.31 0.49	0.141	0.171	0.389
1.2	0.13 0.02 0.04 0.81	0.02 0.13 0.66 0.19	0.05 0.11 0.63 0.21	0.138	0.143	0.664
2	0.13 0.01 0.01 0.85	0.00 0.10 0.81 0.05	0.01 0.14 0.80 0.05	0.139	0.140	0.811

(4.1), the 2 x 2 interaction tables of Z_1, Z_2, and Z_3 and their marginal probabilities are given in Table I. At $\gamma = -1.2$ and $\gamma = 1.2$ the positive and negative association balance out resulting in little or no power for Q^* and Q_{com} to detect the dependence among the Z_js. Similar investigations were done for M = 4 with the same finding.

In summary, the Q statistic is to be preferred for testing independence, it has better power than Q^* or Q_{com}. However, if the interest is on the overall association being zero or that the pairwise associations are all equal, then either Q^* or Q_{com} can be used. Q_{com} is to be preferred since it has been pointed out by Fleiss (1972) that a test statistic of the form Q^* has serious defects if the marginal totals (in our case the n_j's) of the 2 x 2 tables are very different. It can be shown that the statistics Q^* squared and Q_{com} are equal if the n_j's are all equal. That is, if the n_j's are approximately equal then the statistics Q^* squared and Q_{com} are not very different.

4.2 Application to the breeding birds census data

The data is on 5 common species of forest birds on 20 plots of size 6–11 hectares in the eastern deciduous forest (Van Velzen, and Van Velzen, 1988). The data are based on several trips to each plot to determine the presence of the species as a summer resident (territorial). The diagonal elements of Table III are the n_j's (the observed presences for the particular species out of the 20 plots) and the off diagonal elements are

$$\left[N-1\right]^{1/2} \frac{\left[NU_{jk} - n_j n_k\right]}{\left[n_j\left[N-n_j\right]n_k\left[N-n_k\right]\right]^{1/2}},$$

the components of Q and Q^* (equations 3.5 and 3.6). They can be treated as approximate standard normal variates and give informa-

Table II: Breeding birds census data for presence (+) or absence (−) of five species in the eastern deciduous forest.

Census	REV	AF	WT	HW	V
1	+	−	+	−	+
3	−	−	+	−	+
10	+	+	+	+	−
19	+	+	+	−	−
22	+	+	−	−	−
25	+	+	+	+	−
26	+	+	+	+	−
27	+	−	−	−	−
28	+	−	−	−	+
29	+	−	−	−	+
30	+	+	+	+	−
31	+	−	+	+	+
32	+	−	+	−	−
33	+	−	+	−	−
37	+	+	+	+	−
38	+	−	+	−	−
59	+	−	−	−	−
61	+	−	+	−	+
65	+	−	+	−	−
66	+	−	+	−	−

where REV: Red-eyed Vireo, AF: Acadian Flycatcher, WT: Wood Thrush, HW: Hooded Warbler, and V: Veery.

---

tion about the pairwise associations. The data gives the following values for the statistics Q, Q^* and Q_{com}: Q = 20.71, which has P ≐ 0.03, Q^* = 0.36 and P ≐ 0.72, and Q_{com} = 0.33, with P ≐ 0.63. Using the StatXact package, Gajjar *et al* (1989), the exact P-value (two-sided) associated with Q_{com} is P = 0.71. If we assume that the γ_{jk}'s are equal all equal to γ, the maximum likelihood estimates and their standard errors (in paretheses) of α_j's and γ are: $\hat{\gamma}$ = 0.05 (0.09),

	REV	AF	WT	HW	V
$\hat{\alpha}$	1.51	−0.34	0.56	−0.47	−0.47
	(0.52)	(0.24)	(0.26)	(0.26)	(0.26)

Table III: The marginal frequencies (diagonal) and components of Q and Q^* (off diagonal).

	REV	AF	WT	HW	V
REV	19	0.72	-0.58	0.63	-1.53
AF		7	0.79	2.89	-2.09
WT			15	1.65	-0.55
HW				6	-0.83
V					6

Q shows that there is some association among the five species. Examining Table III, it is clear why the P-values are disparate for the three tests. There are negative and positive associations which cancel in Q^* and Q_{com}. We see that species V has a negative association with each of the other species, with a strong association with species REV and AF. Species HW has positive association with species AF and WT.

REFERENCES

BONNEY, G. E. (1987). Logistic regression for dependent binary observations. *Biometrics* **43**, 951-973.

COX, D. R. (1970). Analysis Binary Data. *London: Methuen.*

COX, D. R. (1972). The analysis of multivariate binary data. *Journal of the Royal Statistical Society, Series C,* **21**, 113-120.

EDLEFSEN, L. E., and JONES, S. D. (1987). Gauss version 1.49b,

Aptech Systems, Inc. Kent, WA.

FLEISS, J. L. (1972). Statistical Methods for Rates and Proportions, *Wiley, New York.*

GAJJAR, Y., HILTON, J., MEHTA, C., PATEL, N., SENCHAUDHURI, P., and WALSH, S., (1989). StatXact, Cytel Software Corporation, Cambridge, Massachusetts.

HIRJI, K. F, MEHTA, C. R. and PATEL, N.R. (1988). Exact inference for matched case-control studies. *Biometrics* **44**, 803-814.

HIRJI, K. F, MEHTA, C. R., and TSIATIS, A. A. (1989). Median unbiased estimation for binary data. *American Statistician* **43**, 7-11.

MANTEL, N. (1963). Chi-square tests with one degree of freedom; extensions of the Mantel-Haenszel procedure. *Journal of the American Statistical Association* **58**, 690-700.

MANTEL, N, and Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of diseease. *Journal of the National Cancer Institute* **22**, 719-748.

MEHTA, C. R, PATEL, N. R. and GRAY, R. (1985). Computing an exact confidence interval for the common odds ratio in several 2 x 2 contingency tables. *Journal of the American Statistical associatio* **80**, 969-973.

NERLOVE, M., and PRESS, J. S. (1973). Univariate and multivariate log-linear and logistic models. *R-1306-EDA/NIH, Rand Corporation, Santa Monica, California.*

ROSNER, B. (1984). Multivariate methods in ophthalmology with applications to other paired-data situations. *Biometrics* **40**, 1025-1035.

TRITCHLER, D. (1984). An algorithm for exact logistic regression. *Journal of the American Statistical Association* **79**, 709-711.

VAN VELZEN, A. C., and VAN VELZEN, W. T. (1988). Forty-seventh breeding bird census. *American Birds* **38**, 64-87.