Biometrika (1994), **81**, 2, *pp.* 403–8
Printed in Great Britain

A note on the quadratic exponential binary distribution

BY D. R. COX
Nuffield College, Oxford OX1 1NF, U.K.
AND NANNY WERMUTH
Psychological Institute, University of Mainz, Mainz, D-55099, Germany

SUMMARY

The joint distribution of p binary variables is studied in the quadratic exponential form containing only 'main effects' and 'two-factor interactions' in the log probabilities. Approximate versions of marginalized forms of the distribution are studied based on Taylor expansion and a number of conclusions drawn.

Some key words: Binary data; Log linear model; Multivariate analysis; Yule–Simpson paradox.

1. INTRODUCTION

Consider the joint distribution for p binary random variables $A_1, \ldots, A_p$ in which

$$\log \text{pr}\{A_j = i_j\ (j = 1, \ldots, p)\} = \mu + \sum \alpha_j i_j + \sum_{j>k} \alpha_{jk} i_j i_k. \tag{1}$$

Here $i_j = \pm 1$, μ is a normalizing constant and the α's are unknown parameters. The distribution (1) is sometimes called a quadratic exponential distribution.

In a sense this is the binary analogue of the multivariate normal distribution. Thus it has virtually the same number of unknown parameters and when n independent and identically distributed observations are available the likelihood has the exponential family form with the one-way and two-way tables of frequencies as canonical statistics. Further the conditional distribution of, say, (A_1, A_2) given all remaining variables is formed by subtracting from (1) a function of $i_3, \ldots, i_p$ so that the resulting log odds ratio in the conditional 2×2 tables for (A_1, A_2) is $4\alpha_{12}$, for all values of $(i_3, \ldots, i_p)$. This is qualitatively similar to the interpretation of concentrations, i.e. the off-diagonal elements of the inverse covariance matrix, via the partial correlation of two variables given all remaining variables (Wermuth, 1976).

Similarly, if we condition on all variables except one, say on $A_2, \ldots, A_p$, we have that the conditional logistic transform for A_1 is

$$2\alpha_1 + 2(\alpha_{12} i_2 + \ldots + \alpha_{1p} i_p).$$

Thus $2\alpha_1$ is the average logistic transform, the average being taken over equally weighted values ± 1 for the other variables. Sadly this interpretation has no direct operational force. As is general for log odds ratios, there is a double interpretation of $2\alpha_{jk}$ as the conditional regression coefficient of A_j on A_k and of A_k on A_j, in both cases given all other variables.

Unfortunately, however, the distribution (1) does not retain its exact form under marginalization and conditioning (Cox, 1972). These considerations restrict its usefulness, although Zhao & Prentice (1990) and Fitzmaurice & Laird (1993) have provided methods by which (1) can be combined with a specification of the marginal distribution of individual components.

The object of the present paper is to show that closure under marginalization and conditioning can, however, be achieved approximately; some consequences are explored.

2. EXPANSIONS FOR MARGINAL PROBABILITIES

If we marginalize (1) with respect to A_p, that is compute the distribution of $(A_1, \ldots, A_{p-1})$, we have that

$$\log \text{pr}\{A_j = i_j\ (j = 1, \ldots, p-1)\} = \mu + \log 2 + \sum \alpha_j i_j + \sum_{j>k} \alpha_{jk} i_j i_k + \log \cosh (\alpha_p + \sum \alpha_{jp} i_j), \tag{2}$$

where $\sum$ denotes summation over $j = 1, \ldots, p-1$. Suppose now that the interaction parameters are small.

Because

$$\log \cosh (\alpha + \varepsilon) = \log \cosh \alpha + \varepsilon \tanh \alpha + \tfrac{1}{2}\varepsilon^2 \operatorname{sech}^2 \alpha + \tfrac{1}{6}\varepsilon^3(-2 \sinh \alpha \operatorname{sech}^3 \alpha) + O(\varepsilon^4),$$

it follows that products of three variables, representing departures from (1), arise, as would be expected, first from the cubic term. A typical term is

$$-2\alpha_{j_1 p}\alpha_{j_2 p}\alpha_{j_3 p} \sinh \alpha_p \operatorname{sech}^3 \alpha_p = -\alpha_{j_1 p}\alpha_{j_2 p}\alpha_{j_3 p} g(\alpha_p),$$

say. The function $g(.)$ is odd, is zero at $\alpha_p = 0$ and as $\alpha_p \to \pm\infty$, and is less than 0·8 throughout. In many fields of application a log odds ratio of 1, that is an α_{jk} of $\frac{1}{4}$, would be quite substantial so that in many cases the three-factor interaction will be small and this has been confirmed by numerical work. Note, however, that if p is appreciable there will be many three-factor terms. Note also that if all the α_{jk} and α_p are positive the three-factor contribution is negative.

If we ignore cubic and higher-order terms in the expansion, we have a $(p-1)$-dimensional quadratic exponential model in which the new coefficients are

$$\mu_{(p)} = \mu + \log 2 + \log \cosh \alpha_p + \tfrac{1}{2}\sum \alpha_{jp}^2 \operatorname{sech}^2 \alpha_p, \tag{3}$$

$$\alpha_{(p)j} = \alpha_j + \alpha_{jp} \tanh \alpha_p, \tag{4}$$

$$\alpha_{(p)jk} = \alpha_{jk} + \alpha_{jp}\alpha_{kp} \operatorname{sech}^2 \alpha_p. \tag{5}$$

The argument can now be repeated marginalizing successively over a set of variables, leading to the following general result. Suppose that A is partitioned into two parts, $A = (B, C)$ of dimensions p_B, p_C. Let j, k refer to B, and r, s to C. Then the marginal distribution of B is approximately of the quadratic exponential form with parameters

$$\mu_{(C)} = \mu + p_C \log 2 + \sum_r \log \cosh \alpha_r + \tfrac{1}{2}\sum_{j,r} \alpha_{jr}^2 \operatorname{sech}^2 \alpha_r, \tag{6}$$

$$\alpha_{(C)j} = \alpha_j + \sum_r \alpha_{jr} \tanh \alpha_r + \sum_{r+s} \alpha_{jr}\alpha_{rs} \tanh \alpha_s \operatorname{sech}^2 \alpha_r, \tag{7}$$

$$\alpha_{(C)jk} = \alpha_{jk} + \sum_r \alpha_{jr}\alpha_{kr} \operatorname{sech}^2 \alpha_r. \tag{8}$$

The subscript (C) is a reminder that marginalization over the distribution of C has taken place. The results are best proved inductively starting with the case of three variables.

The conditional distribution of C given B can now be obtained by subtraction of the log probability for the marginal distribution of B from that for the joint distribution of B, C. We obtain another quadratic exponential model to the order of approximation under consideration.

3. SOME IMPLICATIONS

We now consider briefly a number of implications of the above formulae. All the conclusions are, of course, subject to the approximations used in § 2.

First, suppose that, possibly after interchanging the definition of the levels of particular components, we have that $\alpha_{jk} \geqslant 0$ for all j, k. This gives an analogue of the MTP2 multivariate normal distributions (Karlin & Rinott, 1980) in which partial correlations of all orders are nonnegative. It follows from (8) that as we marginalize over more and more components the conditional associ-

ation between any given pair of components is nondecreasing and indeed in most cases is strictly increasing. Further, the Yule–Simpson paradox, a sign change in α_{jk} as the conditioning set changes, is not possible to this order.

Secondly, various special choices of parameters can be made in (1). Thus an analogue of the intraclass model with constant mean would be obtained by setting $\alpha_j = \alpha$, $\alpha_{jk} = \varepsilon$. These are not directly interpretable in terms of marginal properties and two-way associations, although 4ε is the conditional log odds ratio between a pair of variables given all the remaining ones. However (8) shows that the marginal two-way log odds ratio is approximately $4\varepsilon + 4(p-2)\varepsilon^2 \operatorname{sech}^2 \alpha$.

For example, for $p = 4$, Table 1 compares exact and approximate log odds ratios after marginalization; note that $\varepsilon = 0{\cdot}4$ corresponds to a conditional odds ratio of $e^{1{\cdot}6} = 4{\cdot}95$ which in many contexts would be a large association.

Table 1. *Log odds ratios of distributions derived by marginalization from a four-dimensional quadratic exponential distribution with* $\alpha_j = \alpha$, $\alpha_{jk} = \varepsilon$

	$\alpha = 0{\cdot}4$ $\varepsilon = 0{\cdot}4$	$\alpha = 0{\cdot}2$ $\varepsilon = 0{\cdot}2$	$\alpha = 0{\cdot}2$ $\varepsilon = 0{\cdot}4$	$\alpha = 0{\cdot}4$ $\varepsilon = 0{\cdot}2$
		3-dim. distribution		
Exact 1	1·97	0·93	2·05	0·91
Exact 2	2·18	0·96	2·16	0·95
Approx.	2·15	0·95	2·22	0·94
		2-dim. distribution		
Exact	2·81	1·15	3·00	1·10
Approx.	2·70	1·11	2·83	1·07

The values Exact 1, Exact 2 correspond to the conditional 2×2 tables at the two levels of the third variable. The difference between them arises from the small three-factor interaction induced by the marginalization. Approximate values are from (8).

Thirdly, we can adapt the results to provide an approximate partitioning of logistic regression analogous to the partitioning formula for least squares regression which in Yule's notation for regression coefficients can be written

$$\beta_{13} = \beta_{13.2} + \beta_{12.3}\beta_{23}.$$

We consider for simplicity three binary variables. Repeated application of the formulae of § 2 then shows that

$$\gamma_{13} = \gamma_{13.2} + \tfrac{1}{2}\gamma_{12.3}\gamma_{23} \operatorname{sech}^2 \alpha_2,$$

where the γ's are logistic regression coefficients; the factor $\frac{1}{2}$ arises because the levels of the binary variables are coded $(-1, 1)$ rather than $(0, 1)$. For example, $\gamma_{13.2}$ is the logistic regression coefficient of A_1 on A_3 in the regression of A_1 on A_2, A_3, that is taken conditionally on A_2. Note, however, that, because the equation is accurate only to quadratic terms in the regression coefficients, the second term could be written in different forms, for example replacing $\gamma_{12.3}$ by γ_{12}. To this extent, the parallel with the least squares formula is somewhat contrived.

So far we have discussed a single distribution for A, fitting from a simple random sample being via maximum likelihood for a log linear model. If there are explanatory variables one possibility (Fitzmaurice & Laird, 1993) is that the probabilities of each component depend on explanatory variables, for example via a linear logistic model. In this model, which is analogous to a multivariate normal regression, the additional parameters α_{jk} correspond to concentrations in the residual covariance matrix. These second-order parameters would be specified and estimated separately. To express the conditional log odds parameters α_j in terms of the marginal δ_j, we need to invert

the expansion (7). The answer is that to second order

$$\alpha_j = \delta_j - \sum \alpha_{jr} \tanh \delta_r + \tanh \delta_j \sum \alpha_{jr}^2 \operatorname{sech}^2 \delta_r, \tag{9}$$

the summation being for all $r \neq j$. Here δ_r is one-half the marginal log odds for A_r; that is

$$2\delta_r = \log \{\operatorname{pr}(A_r = 1)/\operatorname{pr}(A_r = -1)\}.$$

If now the δ_j are specified by loglinear models, e.g.

$$\delta_j = x_j^{\mathrm{T}} \beta_j,$$

and the α_{jk} left unspecified, or perhaps constrained to be equal, there results a nonlinear model to be fitted by maximum likelihood or, perhaps, in special cases by the device described for so-called reduced models by Cox & Wermuth (1990).

One further possible use of (1) is to set out models for stationary binary time series by setting $\alpha_j = \alpha$ and by re-expressing the second-order parameters in terms of marginal lagged associations. That is, the initial specification (1) of associations would give the log odds of two values a given distance apart conditional on all previous, intermediate and subsequent values and because this is on the whole not a good way to specify and interpret time series structure it has to be replaced by a different specification. We shall not explore this issue here.

Sometimes binary data are produced by median dichotomy of all components. Inversion of (7) into the form (9) shows that this requires that for all j we have $\alpha_j = 0$. Then (8) simplifies to the especially simple form

$$\alpha_{(C)jk} = \alpha_{jk} + \textstyle\sum_r \alpha_{jr}\alpha_{kr}.$$

More detailed calculation shows that in this case the quadratic exponential form is retained to a higher degree of accuracy.

A key property of the multivariate normal distribution is that if, say, A_1 is independent of A_2 and A_1 is independent of A_3 then A_1 is independent of (A_2, A_3). This is true also conditionally on a further set of variables. To the order of accuracy of our expansions, the same property is true for the quadratic exponential distribution. To see this consider $p = 3$ and apply (8) twice to show that to the required order $\alpha_{12} = \alpha_{13} = 0$, thus showing that, for example, A_1 is also conditionally independent of, say, A_2 given A_3, etc. This condition is important for justifying the representation of relations between sets of variables via binary analogues of covariance graphs (Cox & Wermuth, 1993).

Essentially the same algebra shows that if we start, again for three variables, from the assumption that A_1 is conditionally independent of A_2 given A_3 then, to the order being considered,

$$\alpha_{(3)12} = \alpha_{(2)13}\alpha_{(1)23} \operatorname{sech}^2 \alpha_3,$$

in which the α's are measures of dependence of pairs of variables marginalizing over the third variable. This is a fairly direct analogue of the result in normal theory that if Y_1 is conditionally independent of Y_2 given Y_3, then

$$\operatorname{corr}(Y_1, Y_2) = \operatorname{corr}(Y_1, Y_3) \operatorname{corr}(Y_2, Y_3).$$

Finally we consider the analogue in the present context of a single factor model for correlations, namely a latent class model with just two unobserved classes. For this consider p observed binary variables $(A_1, \ldots, A_p)$ as before and an unobserved binary variable L. Suppose that their joint distribution is such that conditionally on L the A's are mutually independent. Then the joint distribution of $(A_1, \ldots, A_p, L)$ is of quadratic exponential form with $\alpha_{jk} = 0$ for $j, k = 1, \ldots, p$.

It follows on marginalizing over L that the joint distribution of $(A_1, \ldots, A_p)$ is approximately of quadratic exponential form with the $\alpha_{(L)jk}$ satisfying a tetrad condition exactly that characterizing the correlations in the single factor multivariate normal model, namely

$$\alpha_{(L)j_1j_2}\alpha_{(L)j_3j_4} = \alpha_{(L)j_1j_3}\alpha_{(L)j_2j_4}.$$

This follows from (8) which leads directly to

$$\alpha_{(L)j_1j_2} = \alpha_{Lj_1}\alpha_{Lj_2}\,\mathrm{sech}^2\,\alpha_L,$$

from which the vanishing of tetrads follows immediately.

4. An empirical example

Finally we give a brief illustration with some empirical data. Table 2 specifies the joint distribution of four binary variables arising in a study (Weck, 1991) of German high school students. The four variables are: A, change of school during high school years; B, repetition of a high school class; C, change of a primary school class; D, father with at least 13 years schooling, the Abitur. For each variable, -1 corresponds to yes, $+1$ to no.

Because of the temporal ordering of the variables, a recursive study of dependence would be reasonable, but here we use the data to examine the joint distribution. Table 2 shows also the fitted frequencies under a quadratic exponential model. The agreement is good, with a likelihood ratio chi-squared value of 6·5 on 5 degrees of freedom.

Table 2. *Distribution of four binary variables concerning German high-school children; fitted frequencies for quadratic exponential model*

A	B	C	D	Obs.	Fitted	A	B	C	D	Obs.	Fitted
−1	−1	−1	−1	40	42·3	−1	−1	−1	+1	22	24·6
+1	−1	−1	−1	49	44·4	+1	−1	−1	+1	44	43·7
−1	+1	−1	−1	44	37·0	−1	+1	−1	+1	16	18·2
+1	+1	−1	−1	100	109·3	+1	+1	−1	+1	95	90·5
−1	−1	+1	−1	77	71·8	−1	−1	+1	+1	78	78·3
+1	−1	+1	−1	125	132·5	+1	−1	+1	+1	247	244·4
−1	+1	+1	−1	62	71·9	−1	+1	+1	+1	71	65·9
+1	+1	+1	−1	385	372·6	+1	+1	+1	+1	571	578·4

Table 3 shows the estimates of the parameters in the quadratic exponential model. To illustrate the formula (8) we computed for each pair of variables the marginal log odds ratio summing out over the complementary pair of variables. These are also shown in Table 3 together with the approximations derived from (8). The agreement in all cases is excellent showing that the marginal properties can indeed be found directly from the coefficients in the quadratic exponential model, i.e. from the conditional log odds ratios.

Table 3. *Data concerning German high-school children: estimated parameters for quadratic exponential model; comparison of marginal log odds and approximation* (8)

	(A, B)	(A, C)	(A, D)	(B, C)	(B, D)	(C, D)
Cond. log odds	1·03	0·56	0·52	0·13	−0·17	0·63
Marg. log odds	1·02	0·67	0·55	0·23	−0·05	0·67
Approx. (8)	1·02	0·68	0·54	0·21	−0·05	0·68

$\hat{\alpha}_A = -0{\cdot}555$, $\hat{\alpha}_B = -0{\cdot}182$, $\hat{\alpha}_C = 0{\cdot}596$, $\hat{\alpha}_D = 0{\cdot}025$;
$\hat{\alpha}_{AB} = 0{\cdot}258$, $\hat{\alpha}_{AC} = 0{\cdot}141$, $\hat{\alpha}_{AD} = 0{\cdot}131$, $\hat{\alpha}_{BC} = 0{\cdot}033$, $\hat{\alpha}_{BD} = -0{\cdot}043$, $\hat{\alpha}_{CD} = 0{\cdot}157$.

References

Cox, D. R. (1972). The analysis of multivariate binary data. *Appl. Statist.* **21**, 113–20.

Cox, D. R. & Wermuth, N. (1990). An approximation to maximum likelihood estimates in reduced models. *Biometrika* **77**, 746–61.

COX, D. R. & WERMUTH, N. (1993). Linear dependencies represented by chain graphs (with discussion). *Statist. Sci.* **8**, 204–83.

FITZMAURICE, G. M. & LAIRD, N. M. (1993). A likelihood-based method for analysing longitudinal binary responses. *Biometrika* **80**, 141–51.

KARLIN, S. & RINOTT, Y. (1980). Classes of orderings of measures and related correlation inequalities I. Multivariate totally positive distributions. *J. Mult. Anal.* **10**, 467–98.

WECK, M. P. (1991). *Der Studienfachwesel. Ein Längsschnittanalyse der Interaktionsstruktur von Bedingungen des Studienver laufs.* Frankfurt: Lang.

WERMUTH, N. (1976). Analogies between multiplicative models in contingency tables and covariance selection. *Biometrics* **32**, 95–108.

ZHAO, L. P. & PRENTICE, R. L. (1990). Correlated binary regression using a quadratic exponential model. *Biometrika* **77**, 642–8.

[*Received July* 1993. *Revised October* 1993]