Collection of Biostatistics Research Archive
COBRA Preprint Series

Year 2010 *Paper* 68

The Linkset Model for 2^n Contingency Tables

Mikel Aickin*

*ErgoLogic Consulting & Software, maickin@earthlink.net

http://biostats.bepress.com/cobra/art68

The Linkset Model for 2^n Contingency Tables

Mikel Aickin

Abstract

Abstract The linkset model is defined for parametrizing the general 2^n contingency table. The linkset parameters are designed to represent latent influences that promote the co-occurrences of binary events beyond that explained by chance. Linkages involving 2 through n binary variables are included in this parametrization. The intent of this process is to elucidate the patterns of linkage, no matter how complex they might be, rather than to fit simplifying models. The relationship between linkset parameters and the natural parameters for a 2n table are derived, and large sample inference methods are provided. Examples are given from medical diagnostics, survival from the Titanic sinking, and employment discrimination in Chicago.

The Linkset Model for 2^n Contingency Tables

Mikel Aickin

Family & Community Medicine
and College of Public Health
University of Arizona

Contact: maickin@earthlink.net, or 4840 N. Valley View Rd., Tucson AZ 85718, USA.

Abstract

The linkset model is defined for parametrizing the general 2^n contingency table. The linkset parameters are designed to represent latent influences that promote the co-occurrences of binary events beyond that explained by chance. Linkages involving 2 through n binary variables are included in this parametrization. The intent of this process is to elucidate the patterns of linkage, no matter how complex they might be, rather than to fit simplifying models. The relationship between linkset parameters and the natural parameters for a 2^n table are derived, and large sample inference methods are provided. Examples are given from medical diagnostics, survival from the Titanic sinking, and employment discrimination in Chicago.

Keywords: binary variables, agreement, diagnosis, Möbius inversion, event co-occurrence

Running title: Linksets for 2^n Contingency Tables

1. Introduction

The analysis of 2^n contingency tables falls somewhere between the special 2×2 case and the more general field of rectangular tables. The 2×2 case has been exhaustively treated in the statistical literature, from a plethora of different viewpoints. In contrast, interest in the general 2^n case has largely revolved around log-linear models or linear models, inspired in part by a few very influential monographs (Bishop, Fienberg, Holland, 1975; Haberman, 1974; Gokhale, Kullback, 1978). With the passage of time application of these models seems to have waned, although logistic regression could be considered a counter-example.

The general approach to contingency table analysis has been in the direction of simplification, either by fitting models that impose restrictions on the cell probabilities, or by focusing on measures derived from the table, such as linear combinations of ln-odds ratios. One reason for this comes from the number of free parameters that are involved. In a Normal distribution of n variables there are $n(n+3)/2$ parameters, but in a distribution over a 2^n contingency table there are 2^n-1, so that parameter reduction seems a worthy aim in and of itself.

The focus of this article is on reparametrizing the 2^n table in such a way that all of the resulting parameters are interpretable, and potentially meaningful, in the context of a model for explaining co-occurrence of events. The linkset model will be defined, and its parameters interpreted. The challenging problem of relating the linkset parameters to the natural parameters on a 2^n table will be resolved, and the asymptotic distribution of estimates will be provided. Practical examples will be given in the areas of surgical diagnostic reliability, survival in the context of the Titanic disaster, and employment discrimination in the United States.

2. The Linkset Model

The 2^n contingency table consists of cells representing all possible combinations of n binary variables (ones that are either 0 or 1, often interpreted as no/yes or absent/present). In the examples below, each such cell will be represented as a string of outcome 0's and 1's, signifying the values of the n-variables in a standard order. Because each cell can be associated with the set of binary variables that "happened" (were 1), it is possible to refer to the 0/1 pattern as a *set*. Thus, for example, 000 would be the empty set (because no variables happened) and 101 would be the set of the first and thrid variables, both of which happened, the second variable not having happened. The *empirical frequency* is the count of the number of cases in which the corresponding specific pattern happened.

The linkset model provides a way of describing how these *observed sets* came about. It posits latent binary variables (*linkset variables*), one for each non-void subset of the variables that define the table. When it is used to refer to the latent variables, rather than the cells in the table, a 0/1 pattern will be called a *linkset*. This is done to avoid confusion, since the latent variables are indexed by the same symbols that are used for the cells in the table.

For example, the linkset variable associated with 101 indicates the event that the first and third variables happened for an underlying reason. This reason is not further defined; the point of the linkset model is to allow such

underlying reasons to be portrayed in the model, irrespective of their character. Note that the table cell 101 can happen in two distinct ways. First, the two variables might have simply chanced to happen together. Secondly, the linkset variable associated with 101 might have happened, causing them to be linked. Likewise, linkset 110 indexes the linkset variable which causes the first two variables to happen for cause (instead of by chance), and the linkset 111 corresponds to all three variables happening for cause. Note especially in this latter case that the *observed event* 111 is completely different from the *linkset* 111, since the former event can happen as a consequence of several different combinations of linkset events (such as 110 and 001, or 110 and 011, and so on). It is the fact that an observed event can happen so many different ways in terms of linkset events that makes the model computations challenging.

The stochastic assumption behind the linkset model is that the latent linkset variables are independent. Each one has a probability in the population, which is denoted by β, and indexed by the corresponding linkset. A tranformation of these parameters, $\lambda n(\beta) = -\ln(1-\beta)$ is important for reasons having to do with both the sampling distribution of the maximum likelihood estimates, and the theoretical computations, as described in the theory sections below, which follow the examples. Note that it is possible for some of the maximum likelihood estimates of the β's to be negative, which will be discussed below.

One of the computations that can be derived from the model is an *attribution probability*. This is the probability that an occurrence in a given cell was due to the linkset variable for that cell, and thus it is the conditional probability of co-occurrence for cause, as opposed to chance. Both the estimates of the β-parameters and the attribution probabilities are given in the examples. Finally, for a given subset of variables, one can compute the probability that they will ever be linked in the table (which could include linkage with other variables. This is called the *link association* and is also used in two of the examples.

3. Example: Diagnostic Reliability

The dataset in Table 1 represents the recommendation for (1) or against (0) carotid endarterectomy on the part of five clinicians, using a standard diagnostic scheme (Uebersax & Grove, 1989). The layout of this table will be used for subsequent examples. The "Set" column gives a 0/1 representation of the binary table cells, while the "Empirical Probability" column gives the sample fraction in each cell. The "Linkset Parameter" columns contains the β-estimates, with the one-sided p-value for the null hypothsis $\beta=0$, and the standard deviation of the estimate $\lambda n\left(\hat{\beta}\right)$. (see the theory sections below). Simulations have found that the asymptotic Normal approximation works better for this transformed parameter. Finally the "Attributable Probability" is the probability that the linkset on the corresponding row was responsible for occurrences in the set on that row.

Table 1. Recommendations for heart surgery by 5 clinicians (n=859).					
Set	Empirical Probability	Linkset Parameter	P-value	SDE	Attributable Probability
00000	.4494	.0000	.	.	.
10000	.0477	.0960	0.000	.0158	.
01000	.0000	.0000	.	.0000	.
00100	.0093	.0203	0.002	.0073	.
00010	.0326	.0676	0.000	.0132	.
00001	.0570	.1126	0.000	.0171	.
11000	.0000	.0000	.	.0000	.
10100	.0093	.0161	0.006	.0065	.894
01100	.0000	.0000	.	.0000	.
10010	.0279	.0439	0.000	.0106	.876
01010	.0000	.0000	.	.0000	.
00110	.0000	-.0014	.	.0006	.
10001	.0489	.0710	0.000	.0133	.876
01001	.0000	.0000	.	.0000	.
00101	.0093	.0155	0.007	.0064	.873
00011	.0058	.0031	0.273	.0051	.289
11100	.0012	.0023	0.159	.0023	.999
11010	.0012	.0021	0.159	.0021	.999
10110	.0047	.0055	0.090	.0041	.683
01110	.0000	.0000	.	.0000	.
11001	.0000	.0000	.	.0000	.
10101	.0268	.0339	0.000	.0089	.832
01101	.0000	.0000	.	.0000	.
10011	.0780	.0921	0.000	.0137	.882
01011	.0000	.0000	.	.0000	.
00111	.0058	.0091	0.024	.0046	.891
11110	.0023	.0036	0.099	.0028	.915
11101	.0047	.0065	0.034	.0036	.931
11011	.0023	.0026	0.127	.0023	.830
10111	.0955	.0903	0.000	.0126	.858
01111	.0000	.0000	.	.0000	.
11111	.0803	.0763	0.000	.0102	.950
Each set represents recommedations for (1) or against (0) surgery by five clnicians. P-values and attributable probabilities are not available for cases where the linkset parameter is 0. SDE = estimate of standard deviation of the estimate of $\lambda n \beta$.					

Due to the nature of the data, one expects linkage among the clinician recommendations, and in fact only the linkage for clinicians 3 and 4 was estimated to be slightly negative. Note that the parameter on the top row is 0 by definition, and consequently the other columns in that row are not meaningful. For pairwise linkages, there is evidence for clinicians (1,3), (1,4), (1,5), (2,5), and (3,5). These would be interpreted as recommendation linkages between these pairs, without being linked to any of the other clinicians. Among the triplets there is evidence for linkage among (1,2,3), (1,3,5), (1,4,5), and (3,4,5), with a similar interpretation. The apparently significant quartets are (1,2,3,5), and (1,3,4,5). The quintet of all five is also significant.

The linkset parameters in this example can be considered to pertain to the population of patients considered for heart surgery. The singleton linkset probabilities indicate the fraction of the population that will be recommended for surgery by each clinician, without linkage with the other clinicians. The value of 0 for clinician 2 is odd, in that it indicates no willingness to recommend for surgery without some kind of linkage with the other clinicians. Of the multiple linkage probabilities, that pertaining to (1,2,3,4,5) stand out (8% of the population), along with (1,4,5) at 9%, and (1,5) at 7%.

It may be noted in this example that not all possible combinations of recommendations appear. For this anaysis it was taken that a zero empirical frequency was a "chance zero", rather than a "structural zero" which would have implied that the cell was not logically possible.

Part of the intent of this example is to provide an assessment of the five clinicians. Clinician 1 appears in 9 of the 12 significant linksets, suggesting that perhaps this person employs information or guidance from the individual cases which is used in some shared manner by the others. Indeed, clinician 1 is in both of the significant quartets, the other quartet in which that clinician appears has marginal p-value 0.10, and the quartet in which he/she plays no role never occurs. Another potentially interesting finding is that clinicians 2 and 4 appear never to be linked (except in the (1,2,3,4,5) linkset).

Another important use of this dataset is to estimate the reliability of the surgery recommendation procedure, at least for this sample of five clinicians. We may interpret the attributable probability as a measure of reliability, in the sense that it is the fraction of times, when the clinicians agree, that they do so for some underlying reason, hopefully based on the patient's record and the recommendation guidelines, rather than simply by chance. Here the reliability of a 5-clinician recommendation is 0.95. Since this comprises such a small fraction of cases, it is of interest to note that reliability is also very high for all 4-clinician recommendations, and most of the 3-clinician recommendations, with the possible exception of (1,3,4). Perhaps somewhat surprisingly, the reliability remains high for 2-clinician recommendations, with the exception of (4,5). Note that there is no information in Table 1 concerning the reliability of recommendations against surgery. For this we simply reverse the coding of all recommendations, so that 1 becomes a recommendation against rather than for. Instead of repeating the entire analysis, the reliability results are collected in Table 2. The reliabilities appear uniformly high and relatively free of the number of clinicians recommending against, with the possible exceptions of (2,5) and (1,2,3). Interestingly (2,3,4,5) has only medium reliability, and while this might suggest that the agreement of clinician 1 is important for a negative recommendation, there is a mixed message from the 3-clinican cases in which he/she recommends against. It is also odd that (1,2) has perfect reliability, whereas (1,2,3) is second worst. This may be a hint that there are two or more potentially ambiguous aspects to the surgery recommendation guidelines, although if there is it certainly does not show up uniformly.

Table 2. Reliability of recommendations against heart surgery	
Set	Attributable Probability
11000	1.000
01100	.965
01010	.793
01001	.406
00101	.942
00011	.884
11100	.543
11010	.888
01110	.716
01101	.805
01011	.719
11110	.861
11101	.961
11011	.895
01111	.647
11111	.880
Sets in this table are complements of the sets in Table 1.	

4. Example: Survival on the Titanic

Two-thirds of the people on the RMS Titanic died when it struck an iceberg and sank in April 1912. The loss of life was charged mainly to the lack of lifeboats, but stories from the survivors also suggested that the available lifeboats were not fully utilized, due to a "women and children first" ethic, and also to issues involving the different passenger classes. Table 3 shows a linkset analysis of the factors *survived*, *passenger* (as opposed to crew), and *female*. (Children are not included in the analyses here.)

Table 3. Survival on the Titanic, Adults Only (n=2092).						
Set	Empirical Probability	Linkset Parameter	P-value	SDE	Attributable Probability	Link Association
000	.320	0	.	.	.	.
100	.092	.223	0.0000	.018	.	.115
010	.315	.496	0.0000	.027	.	.161
001	.001	.004	0.0416	.003	.	.200
110	.070	-.026	.	.012	.	.096
101	.010	.022	0.0000	.005	.957	.137
011	.051	.072	0.0000	.008	.972	.181
111	.141	.119	0.0000	.009	.839	.119
Sets are in order: survival, passenger, female. SDE = estimate of standard deviation of the estimate of $\lambda_{n\beta}$.						

Largest linkage (12%) occurs among all three factors, and there it appears that they were linked in 84% of cases. Survival was also linked to *female* (2%) and in these cases 96% was due to linkage. The link between *female* and *passenger* is due to there being relatively few women among the crew. The negative linkage between *survival* and

passenger is hard to explain, and has not been advanced in the historical record as a hypothesis. The total linkage (link association) between *survived* and *passenger* was 10%, and was only slightly higher (14%) between *survived* and *female*.

There are several different ways to look at survival among the passengers, and here we just take one view in Table 4. The largest linkset parameter is 0.36 linking *died* and *male*, with 99% of such cases being linked. This provides substantial evidence for the contention that there was a bias in favor of rescuing the women first. The second most notable linkage is between *died* and *third-class*, substantiating part of the assertions that have been made about third-class passengers not being given the same access to lifeboats. Note that the linkage between *died* and *second-class* is significant, if somewhat modest. But in both cases the attributable probability is very high. Finally, note that both the gender and class stories are bolstered by the triple-linkages of *died*, *male*, *second-class*, and *died*, *male*, *third-class*. The appreciable negative linkage between *male* and *second-class* reflects the composition of the passenger list, as does the positive linkage between *male* and *third-class*.

Table 4. Death on the Titanic, Adult Passengers Only (n=1207)

Set	Empirical Probability	Linkset Parameter	P-value	SDE	Attributable Probability	Link Association
0000	.115	0	.	.	.	.
1000	.003	.0277	0.0227	.0140	.	.693
0100	.047	.2893	0.0000	.0454	.	.584
0010	.066	.3636	0.0000	.0509	.	.136
0001	.062	.3518	0.0000	.0501	.	.480
1100	.097	.3648	0.0000	.0431	.9861	.552
1010	.010	.0452	0.0025	.0164	.8241	.188
0110	.011	−.0638	.	.0233	.	.095
1001	.073	.2809	0.0000	.0365	.9756	.404
0101	.062	.1266	0.0000	.0338	.5874	.276
1110	.127	.1490	0.0000	.0315	.5615	.149
1101	.320	.1715	0.0000	.0433	.4192	.172

Sets are in the order: died, male, second class, third class. SDE = estimate of standard deviation of the estimate of $\lambda n\beta$.

Table 4 is an example of a situation in which the total linkage probabilities (link association) are quite a bit higher than the individual linkset probabilities. Thus *died* and *male* were linked in 55% of the sample, *died* and *second-class* were linked in 19%, and *died* and *third-class* were linked in 40%. Thus, substantial fractions of the sample were linked in ways suggested by the historical record.

5. Example: Employment Discrimination

The data in Table 5 are from an innovative study of labor market discrimination (Bertrand & Mullainathan 2004). The authors created fictional resumés which they then sent out in response to requests for job applications. For the data used here, the characteristics of the resumés were *accepted* (meaning that there was follow-up by the job offeror), *white*, *male*, and *high-quality*, meaning indications of better experience. Gender and ethnicity were indicated by the names of the pseudo-applicants. The full sample had Chicago and non-Chicago subsamples, but here we use just the Chicago subsample.

Table 5. Job follow-up, gender, ethnicity, and quality. (n=2704)

Set	Empirical Probability	Linkset Parameter	P-value	SDE	Link Association
0000	.2067	0	.	.	.
1000	.0096	.0444	0.0000	.0089	.023
0100	.2026	.4950	0.0000	.0297	.009
0010	.0273	.1169	0.0000	.0144	.011
0001	.2085	.5022	0.0000	.0299	.004
1100	.0144	.0115	0.0493	.0070	.014
1010	.0040	.0112	0.0135	.0051	.004
0110	.0273	.0011	0.4547	.0102	.000
1001	.0122	.0057	0.1954	.0066	.004
0101	.2004	−.0049	.	.0210	.
0011	.0303	.0057	0.2912	.0104	.001
1110	.0033	−.0028	.	.0034	.000
1101	.0184	.0019	0.3569	.0052	.005
1011	.0011	−.0070	.	.0029	.
0111	.0292	−.0005	.	.0073	.
1111	.0040	.0035	0.0649	.0023	.003

Sets are ordered: accepted, white, male, high-quality. SDE = estimate of standard deviation of the estimate of $\lambda_n\beta$.

The linksets that actually link events have rather small probabilities, even though two of them are formally significant; *accepted* and *white*, and *accepted* and *male*. Given the expectations that one might have had for this experiment, and the relatively large sample size, it is noteworthy how small the linkages are. On the supposition that individual linkset probabilities might be hiding an overall linkage effect, the link association probabilities are given in Table 5. Each probability here is the fraction of the population that has the specific linkset factors linked, but including linkages with other factors as well. The results, however, substantiate the message of the linkset probabilities themselves, that all linkages are very small.

6. Statistical Theory

There are two complementary ways to look at a 2^n contingency table. The first employs a random vector $(x_1,\ldots,x_n)$ of binary (or Boolean) variables. The sample space in this case is the set of all possible n-vectors of 0's and 1's. The second approach focuses on the random set of subscripts for which the corresponding x's happened, $S = \{i : x_i = 1\}$. In this case the sample space is the set of all subsets of $N = \{1,2,\ldots,n\}$. Since $x_i = 1$ if and only if $i \in S$, the two viewpoints are equivalent. The second approach is more useful for defining the linkset model.

Let $\mathcal{S}$ denote a subset of the collection of all subsets of N. For each nonvoid $L \in \mathcal{S}$ let b(L) denote a binary random variable, such that $\{b(L) : L \in \mathcal{S}\}$ are independent. The subsets L are called *linksets*, for reasons that will become apparent. Further define the set function β on $\mathcal{S}$ by $E[b(L)] = \beta(L)$. We define $\beta(\varnothing) = 0$ for completeness. These values are the *linkset parameters*. The linkset model can then be very simply defined in terms of the random set of subscripts S by

$$S = \bigcup \{L \in \mathcal{S} : b(L) = 1\}$$

Two tasks must be carried out to use this model effectively. The first is to understand the meaning of the linkset parameters, and this will be done with unitary algebra. The second is to determine a 1-1 correspondence between the linkset parameters and the conventional table probabilities, and to derive large sample estimates, and this will be done by employing elements of the theory of random variables taking values in partially ordered sets.

6.1 Unitary Algebra

Unitary algebra is designed to facilitate probability computations for binary random variables. For u and v in the unit interval [0,1] (*unitary numbers*), the definitions are

$u \vee v = u + v - uv$

$u^* = 1-u$

$u \backslash v = (u-v)/v^*$ (=0 if v=1)

$\lambda n(u) = -\ln(u^*)$

$\varepsilon xp(r) = 1 - \exp(-r)$ $(r \geq 0)$

The following basic facts are easily shown

(1) $\vee$ is commutative and associative; $u \vee 0 = u$, $u \vee 1 = 1$

(2) $\lambda n(u \vee v) = \lambda n(u) + \lambda n(v)$

(3) εxp is the inverse of λn, and $\varepsilon xp(r+s) = \varepsilon xp(r) \vee \varepsilon xp(s)$

(4) $(u \vee v)^* = u^* v^*$

(5) $(u \backslash v)^* = u^*/v^*$ $(v<1)$

(1) makes it possible to define the $\vee$ operation over finite collections, and then (2)-(4) can be extended to the $\vee$-sum of any finite collection. It is obvious that $\vee$ is the Boolean "or" operation when restricted to binary variables. When u and v are independent unitary variables, then $E[u \vee v] = E[u] \vee E[v]$, which indicates why one wants the Boolean "or" to be extended to unitary numbers. This result also extends to $\vee$-sums of independent unitary variables. The \

operator corresponds to subtraction, since $(u \vee v) \backslash v = u$. Sheps was the first to understand the importance of this operation for comparing probabilities; $u \backslash v$ is her measure of excess occurrence (Sheps, 1959). Finally, (4) is DeMorgan's rule and (5) is another version of DeMorgan's rule.

The connection between unitary algebra and the linkset model is

$$x_i = \vee\{b(L) : i \in L\}$$

For a pair of variables,

$$x_i x_j = \vee\{b(I)b(J) : i \in I \setminus J, j \in J \setminus I\} \vee \vee\{b(L) : i, j \in L\}$$

In this expression, each b(I)b(J) indicates an event from which x_i happens and x_j happens, but their co-occurrence is taken to be due to chance. Thus, the first $\vee$-sum consists of all events where this chance co-occurrence happens. The second $\vee$-sum consists of all events indicated by b(L) in which both i and j are in the linkset L, and thus their co-occurrence is taken to be linked. This shows the sense in which the occurrence of an indicator like b(L) causes a linked occurrence of all the x's with subscripts in L. The linkset model itself does not define what this linking relationship is, it simply provides a set of latent variables (the b's) which can account for it.

6.2 Identifiability and Estimation of Linkset Parameters

The basic analytic problem with the linkset model is that the cell probabilities become very complicated expressions of the β-parameters. In order to derive relatively simple relations, we turn to the theory of partially ordered sets. First, let π denote the probability distribution of the set S of random subscripts:

$$\pi(T) = P[S = T] \qquad (T \in \mathcal{S})$$

This is an example of a set function on $\mathcal{S}$. The *cumulant* of π (or any set function on $\mathcal{S}$) is

$$\mathbf{C}\pi(T) = \sum\{\pi(U) : U \subseteq T, U \in \mathcal{S}\}$$

We take this to be defined on $\mathcal{S}$. To simplify the computations we assume that $\varnothing \in \mathcal{S}$. In effect, we must also assume $\pi(\varnothing) > 0$, since otherwise

$$\pi(\varnothing)^* = \vee\{\beta(U) : U \in \mathcal{S}\} = 1$$

would imply that some $\beta(U) = 1$, trivializing the model.

The first step toward identifiability is to establish a relationship between the set functions π and β. For any fixed T we have

$$P[S \subseteq T] = \mathbf{C}\pi(T)$$

But also

$$[S \subseteq T] = \prod \{b(U)^* : U \cap T^* \neq \varnothing, U \in \mathcal{S}\}$$

where T* is the complement of T in N. Consequently

$$-\ln \mathbf{C}\pi(T) = \sum \{\lambda n(\beta(U)) : U \cap T^* \neq \varnothing, U \in \mathcal{S}\} = \mathbf{C}\lambda n\beta(N) - \mathbf{C}\lambda n\beta(T)$$

where we write $\lambda n\beta$ for the set function $U \to \lambda n(\beta(U))$ and $\mathbf{C}\lambda n\beta$ for its cumulant. (We also interpret $C\lambda n\beta(N)$ to be the cumulant over all sets in $\mathcal{S}$ in case N is not in $\mathcal{S}$.) Note the special case

$$-\ln \pi(\varnothing) = \mathbf{C}\lambda n\beta(N)$$

These computations give us the complementary results

$$\mathbf{C}\pi(T) = \exp(\mathbf{C}\lambda n\beta(T) - \mathbf{C}\lambda n\beta(N))$$
$$\mathbf{C}\lambda n\beta(T) = \ln \mathbf{C}\pi(T) - \ln \pi(\varnothing)$$

establishing a 1-1 relationship between the cumulants of π and $\lambda n\beta$.

For the final identifiability step we use the fact that $\subseteq$ is a partial order on $\mathcal{S}$. The Möbius function $\mu(V,U)$ is a function of pairs V, U in $\mathcal{S}$, such that for any set function α on $\mathcal{S}$ if we define

$$\mathbf{M}\alpha(U) = \sum \{\alpha(V)\mu(V,U) : V \subseteq U; V \in \mathcal{S}\}$$

then $\mathbf{MC}\alpha = \alpha$ (Rota, 1964). Since the Möbius function exists for any finite partially ordered set, we can apply it to both sides of the previous two equations, obtaining π as a function of β and $\lambda n\beta$ as a function of π, thus also β as a function of π:

$$\pi(T) = \exp(-\mathbf{C}\lambda n\beta(N))\mathbf{M}(\exp \mathbf{C}\lambda n\beta)(T)$$
$$\beta(T) = \varepsilon\exp(\mathbf{M}\ln \mathbf{C}\pi(T)) \qquad (T \neq \varnothing; \beta(\varnothing) = 0)$$

It is worth noting in the special case when $\mathcal{S}$ consists of all subsets of N, that $\mu(V,U) = (-1)^{\#U-\#V}$ when $V \subseteq U$ and 0 otherwise, where #U is the number of elements in U. The Möbius function can be computed in a simple sequence of iterative steps in the general case.

We take the usual empirical probabilities $\hat{\pi}(T)$ to estimate π, and then use the above formulas to produce the estimates $\hat{\beta}(T)$. For inference it turns out to be both easier and more accurate to work with $\lambda n\hat{\beta}(T)$. Directly from the formulas,

$$\mathrm{var}\left(\lambda n\hat{\beta}(T)\right) = \sum \{\mathrm{cov}(\ln \mathbf{C}\hat{\pi}(U), \ln \mathbf{C}\hat{\pi}(V))\mu(U,T)\mu(V,T) : U, V \in \mathcal{S} : U \subseteq T, V \subseteq T\}$$

Applying the delta method to the known form of the covariances of multinomial estimates

$$\mathrm{cov}(\ln \mathbf{C}\hat{\pi}(U), \ln \mathbf{C}\hat{\pi}(V)) \cong \frac{\mathrm{cov}(\mathbf{C}\hat{\pi}(U), \mathbf{C}\hat{\pi}(V))}{\mathbf{C}\hat{\pi}(U)\mathbf{C}\pi(V)} = \frac{\mathbf{C}\hat{\pi}(U \cap V) - \mathbf{C}\hat{\pi}(U)\mathbf{C}\hat{\pi}(V)}{n\mathbf{C}\hat{\pi}(U)\mathbf{C}\hat{\pi}(V)}$$

from which it follows from properties of the Möbius function that

$$\mathrm{var}\left(\lambda n\hat{\beta}(T)\right) \cong \frac{1}{n}\sum \left\{ \frac{C\hat{\pi}(U \cap V)}{C\hat{\pi}(U)C\hat{\pi}(V)} \mu(U,T)\mu(V,T) : U, V \in \mathcal{S} : U \subseteq T, V \subseteq T \right\}$$

where n is the sample size. The validity of this approximation has been confirmed in simulations.

Estimates of the $\beta(T)$ parameters will be presented in examples below, but first it is useful to point out some derived quantities of interest. Note that $\beta(T)$ represents the fraction of the population in which all of the subscripts in T represent mutually linked events, none of which are linked with any subscripts outside of T. If one wants an overall measure of the degree of linkage of variables in T irrespective of linkage to other variables, then it is given by the *link association probability*

$$\overline{\beta}(T) = \vee\{\beta(U) : T \subseteq U, U \in \mathcal{S}\}$$

When T is a singleton, {i}, then $\subseteq$ is replaced in the definition by $\subset$, since b({i}) does not link any events.

The second quantity of interest concerns the fraction of observed cases of S=T which can be attributed to the link indicator b(T). This is the *attributable probability* $\tilde{\beta}(T) = P[b(T) = 1 | S = T]$, and it is a measure of the degree to which the co-occurrences represented by T are due to T-defined linkage. Elementary considerations give

$$\tilde{\beta}(T) = \frac{P[b(T) = 1, S \subseteq T]}{P[S = T]} = \frac{\beta(T) \mathbf{C} \pi(T)}{\pi(T)}$$

The identifiability derivation, presented above, establishes a 1-1 relationship between all probability distributions on a 2^n table and a space of vectors of β-parameters. (The cases in which $\pi(\varnothing)=0$ are present only in the limit.) The linkset model has been motivated in terms of positive β's, representing probabilities of linkage among subsets of events. The general mathematical result does not require positivity, however, and so by permitting negative β's we have an alternative parametrization of the complete set of table distributions. In cases where some of the estimated β(L)'s are negative, we would associate the corresponding linkset L with an "unlinking" or suppression of linkage. It is not yet possible to provide a structural model for suppression on the same footing as for linkage. For this reason, in the examples we note instances of apparent unlinking but do not attempt any inference for the corresponding negative β-parameters.

6.3 The 2×2 Table

The only linkset example that is easy to see is the 2^2 table. A conventional parametrization is shown on the left in Table 6, and the linkset parametrization is shown on the right.

Table 6. Conventional and linkset parametrizations for a 2×2 table; cell probabilities and marginal probabilities.

	Column = 0	Column = 1			Column = 0	Column = 1	
Row = 0	p_{00}	p_{01}	p_{0+}	=	$\beta_1^*\beta_2^*\beta_{12}^*$	$\beta_1^*\beta_2\beta_{12}^*$	$\beta_1^*\beta_{12}^*$
Row = 1	p_{10}	p_{11}	p_{1+}		$\beta_1\beta_2^*\beta_{12}^*$	$\beta_1\beta_2 \vee \beta_{12}$	$\beta_1 \vee \beta_{12}$
	p_{+0}	p_{+1}			$\beta_2^*\beta_{12}^*$	$\beta_2 \vee \beta_{12}$	

β-parameters are shown with subscripts; for example, $\beta_{12} = \beta(\{1,2\})$

It is evident that $\beta_{12} = 0$ is equivalent to independence. It is also easy to express the p-parameters in terms of the β parameters and *vcie versa.* An interpretation of β_{12} emerges from consideration of the extent to which the column variable influences the row variable. To see this, first define

$$\alpha_c = P[\text{Row} = 1 | \text{Col} = c] \qquad c = 0,1$$

We can then compute for Sheps measure of excess occurrence (of row 1 in column 1) that

$$\alpha_1 \setminus \alpha_0 = \frac{\beta_{12}}{\beta_2 \vee \beta_{12}}$$

Sheps defined another measure of excess occurrence, the excess of row 0 in column 0, as $\alpha_0^* \setminus \alpha_1^*$. Her first measure was designed to capture the causal effect of [Col=1] on [Row=1], whereas the second was supposed to capture the dual causal effect of [Col=0] on [Row=0]. In epidemiologic terms the first is a measure of the column variable as a risk factor for the row variable, and the second is a measure of the absence of the column variable as a preventive factor for the row variable. One can combine them with $\vee$ in order to get an overall measure of association, capturing both the risk and prevention effects, and then using unitary algebra we find

$$(\alpha_1 \setminus \alpha_0) \vee (\alpha_0^* \setminus \alpha_1^*) = \frac{\beta_{12}}{\beta_1\beta_2 \vee \beta_{12}}$$

It is of some interest that the expression on the right is (OR-1)/OR, an increasing transform of the odds ratio OR in the original table. Thus the epidemiologic odds ratio does not measure a risk effect, as usually claimed, but instead it measures a combination of risk and prevention effects.

Simiarly easy computations show that

$$P[\text{Row} = \text{Col} : \beta_1, \beta_2, \beta_{12}] \setminus P[\text{Row} = \text{Col} : \beta_1, \beta_2, \beta_{12} = 0] = \beta_{12}$$

Thus β_{12} is the parametric form for Cohen's kappa (Cohen, 1960), a measure of chance-corrected agreement.

7. Other Modeling Approaches

The linkset model has some similarities with a few other approaches in the literature, but most proposed analyses for 2^n tables differ from linksets in fundamental ways. In one of the earliest paper devoted to the 2^n table, Bahadur 1961) derived the consequences of invariance of the joint distribution with regard to certain sets of permutations, with subsequent implications about marginal sums having binomial distributions. This serves as an example of the approach based on imposing simplifying conditions on the table probabilities. From the 1970's onward, perhaps the most frequent approach has been based on increasingly complicated linear combinations of ln-odds ratios, which do not appear to have any natural relationship to linksets. The literature on log-linear and related models is enormous, and is not reviewed here.

With regard to linkage suppression, Boyles & Samaniego (1984) have used the concept of a set of "shocks" to reduce the natural occurrence of 1's in the table. They restrict consideration to "positively dependent" variables, which amounts to $\mathbf{C}\pi(T_1 \cup T_2) \geq \mathbf{C}\pi(T_1)\mathbf{C}\pi(T_2)$, and which is evidently more stringent than the linkset model, even with all β-parameters positive. Presumbly the Boyles & Samaniego metaphor could be developed for linkage suppression, to account for negative linkset parameters.

Many of the restricted-parameter multivariate binary models involve discriminant analysis (Goldstein, 1977), clustering (Martin & Bradley, 1972; Ott & Kronmal, 1976), or the estimation or prediction of an outcome based on an array of binary factors. One path in this latter direction is called "logic regression", in which the binary variables are repackaged into Boolean combinations that are optimal for the model under study (Ruczinski, Kooperberg, LeBlanc, 2003). The focus is thus on how to re-express the binary factors so as to clarify their capacity to predict or explain outcomes. In linkset analysis, the emphasis is not so much on the outcome-predictor dichotomy as it is on explaining the co-occurrence of events.

The reliability and test theory literature on the 2^n table is also very large and is not reviewed here. In general these methods either rely on explanatory continuous latent variables (item response theory), or are oriented toward establishing the reliability or validity of binary tests. Test theory concentrates on the development of item batteries, and although this does involve multivariate binary data, the focus is not so much on understanding the interrelationships among test items as it is to produce a final test with good characteristics. Studies of diagnostic procedures focus on the diagnostic procedure itself, with the aim of establishing its reliability and validity. Seldom does one see cases in which assessing the raters is the point of the research (as in the surgical diagnostic example given here).

Boolean factor analysis attempts to express the relationship between objects and attributes in terms of an objects-factors relationship and a factors-attributes relationship, with the factors being chosen as parsimoniously as possible (Belohlavek & Vychodil, 2007). This approach is like linksets in that it attempts to explain co-occurrences, but unlike linksets in that its goal is to find simple structure rather than to display the structure that is present, irrespective of whether it is simple.

Although it seemed to play a small role above, the Möbius inversion theorem of Rota is actually central to the linkset approach. Despite its technical simplicity, this method has appeared only rarely in the statistical literature, and is usually involved with the theory of random sets. All of the theory of log-linear models, and the more general theory of reconstruction can be framed in terms of the incidence algebra of partially ordered sets and Möbius inversion.

8. Discussion

The linkset model is a latent factor model, and like all such formulations it attempts to provide a way of doing inference about forces that are below the level of direct observation. While it may be impossible to prove the existence of the underlying concepts using a latent factor model, it does seem worthwhile to posit that the concepts exist, and then it makes sense to employ models that explicitly incorporate them into the analysis.

In some cases it is possible to speculate usefully about why events might co-occur, even if the reasons for co-occurrence remain to some extent obscure. In the surgery recommendation example it is plausible that the clinicians are all trained in the same system, understand the rationale for the guidelines, and see how to apply them to different patients. Thus, the reasons for expecting linkages are clear, based on common training. In the Titanic example historical accounts suggested what might link gender or passenger class with survival, and again this does not seem particularly difficult to understand. The employment discrimination example is interesting because of the massive evidence of discrimination against blacks and women in the United States, and yet this clever experiment failed to find it. In the only previous publication using linksets (Aickin & Taetle, 2006) the method was employed to detect co-occurrence of genetic abnormalities in ovarian tumors, which might be related to survival. Here the underlying linking mechanism is less clear, but would presumably be based on the argument that genetic abnormalities play a role in pathways that regulate tumor cell behavior, so that particular co-occurrences would signal the activation of particular pathways. This argument generalizes to genomic and proteomic studies in both diseased and healthy tissues.

On the statistical side, the 2^n contingency table has been somewhat of an orphan. Epidemiologists break such tables down into a series of 2×2 tables that cannot possibly reconstruct the linkages, except perhaps in cases where they involve only pairs of factors. Almost all biostatistical approaches to the 2^n table try to reduce the size of the parameter space in some way, in the search for structure. This enterprise takes it for given that the natural parametrization of the probability distributions on a 2^n table does not reveal its structure. It also raises issues about the assumptions that underly the simplifying models, which is then frustrated by the usual difficulty in validating the assumed conditions. The attempt here has been not so much to simplify the distribution on the 2^n table, but instead to provide an alternative parametrization, in such a way that each parameter says something about structure. The definition of structure here is linkage, meaning that events co-occur for some reason, rather than through simple random conjunction.

Software for performing linkset analyses is available from the author at www.ergologic.us.

Acknowledgment: this research was supported by grant CA094750 from the National Cancer Institute, National Institutes of Health.

References

Aickin M, Taetle R. (2006) Linksets of tumor chromosome breakpoints related to survival in ovarian adenocarcinoma. *Cancer Genetics and Cytogenetics* **166(1)**, 22-26

Bahadur RR. (1981) A representation of the joint distribution of responses to n dichotomous items. In Solomon H (ed.) Studies in Item Analysis and Prediction. Stanford CA: Stanford University Press, 158-168,

Belohlavek R, Vychodil V. (2007) Formal concepts as optimal factors in boolean factor analysis: implications and experiments. In Fifth International Conference on Concept Lattices and Their Applications.

Bertrand M, Mullainathan S. (2004) Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *The American Economic Review* **94(4)**, 991-1013

Bishop YMM, Fienberg SE, Holland PW. (1975) Discrete Multivariate Analysis: Theory and Practice. Cambridge MA: The MIT Press.

Boyles R, Samaniego FJ. (1984) Modeling and inference for multivariate binary data with positive dependence. *Journal of the American Statistical Association* **79**, 188-193

Cohen J. (1960) A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* **20**,37-46

Gokhale DV, Kullback S. (1978) The Information in Contingency Tables. New York NY: Marcel Dekker.

Goldstein M. (1977) A two-group classification procedure for multivariate dichotomous responses. *The Journal of Multivariate Behavioral Research* **12**, 335-346

Haberman S. (1974) The Analysis of Frequency Data. Chicago IL: University of Chicago Press.

Martin DC, Bradley RA. (1972) Probability models, estimation, and classification for multivariate dichotomous populations. *Biometrics* **28(1)**, 203-221

Ott J, Kronmal RA. (1976) Some classification procedures for multivariate binary data using orthogonal functions. *Journal of the American Statistical Association* **71**, 391-399

Rota G-C. (1964) On the foundations of combinatorial theory I. Theory of Möbius functions. *Zeitschrift für Wahrscheinlichkeitstheorie* **2**, 340-368

Ruczinski I, Kooperberg C, LeBlanc M. (2003) Logic regression. *Journal of Computational and Graphical Statistics* **12(3**), 475-511

Sheps MC. (1959) An examination of some methods of comparing several rates or proportions. *Biometrics* **15,** 87-97

Uebersax J, Grove W. (1989) Latent structure agreement analysis. Santa Monica CA: The RAND Corporation,.