STATISTICS IN MEDICINE, VOL. 14, 2143–2160 (1995)

EXACT LOGISTIC REGRESSION: THEORY AND EXAMPLES

CYRUS R. MEHTA AND NITIN R. PATEL

Department of Biostatistics, Harvard School of Public Health, U.S.A., and Cytel Software Corporation, 675 Massachusetts Ave., Cambridge, MA 02139, U.S.A.

SUMMARY

We provide an alternative to the maximum likelihood method for making inferences about the parameters of the logistic regression model. The method is based appropriate permutational distributions of sufficient statistics. It is useful for analysing small or unbalanced binary data with covariates. It also applies to small-sample clustered binary data. We illustrate the method by analysing several biomedical data sets.

1. INTRODUCTION

This paper deals with exact conditional inference for the parameters of the logistic regression model that describes the relationship between a dichotomous outcome and a set of explanatory variables. It is customary to maximize the unconditional likelihood function for parameter estimation, and to perform hypothesis tests with either the Wald, the likelihood ratio, or the efficient scores statistics. For data sets with small sample sizes or unbalanced structure, and for highly stratified data, these asymptotic methods are unreliable. An alternative approach is to base the inference on exact permutational distributions of the sufficient statistics that correspond to the regression parameters of interest, conditional on fixing the sufficient statistics of the remaining parameters at their observed values. This approach, suggested by Cox,[1] was not considered computationally feasible until the development of fast algorithms for deriving these distributions in work by Tritchler,[2] Hirji *et al.*,[3,4] and Hirji.[5] Breslow and Day[6] presented a related asymptotic conditional approach for logistic regression on matched sets. These investigators proposed treating each matched set as a separate stratum and eliminating all stratum-specific parameters from the likelihood function by conditioning on their sufficient statistics. The inference is then based on maximizing a conditional likelihood function. Although easier, computationally, than the exact permutational approach, conditional maximum likelihood estimation is not a trivial problem. Gail *et al.*[7] developed a recursive algorithm to do the computations efficiently.

This paper describes the underlying theory for exact conditional inference, summarizes recent algorithmic developments that make this type of inference computationally feasible, and provides several illustrative examples that contrast exact conditional inference with the more customary unconditional maximum likelihood approach.

2. MODELS, LIKELIHOOD AND SUFFICIENT STATISTICS

We consider two classes of models: logistic regression for unstratified binary data, and logistic regression for stratified binary data. In this section we discuss a uniform method of exact inference for both models, based on permutational distributions of appropriate sufficient statistics.

CCC 0277–6715/95/192143–18

Received July 1993
Revised November 1994

2.1. Logistic regression for unstratified binary data

Consider a set of independent binary random variables, $Y_1, Y_2, \ldots, Y_n$. Corresponding to each random variables, Y_j, there is a $(p \times 1)$ vector $\mathbf{x}_j = (x_{1j}, x_{2j}, \ldots, x_{pj})'$ of explanatory variables (or covariates). Let π_j be the probability that $Y_j = 1$. Logistic regression models the dependency of π_j on $\mathbf{x}_j$ through the relationship

$$\log\left(\frac{\pi_j}{1-\pi_j}\right) = \gamma + \mathbf{x}_j'\beta, \tag{1}$$

where γ and $\beta \equiv (\beta_1, \beta_2, \ldots, \beta_p)'$ are unknown parameters. The likelihood function, or probability of an observed set of values, $y_1, y_2, \ldots, y_n$, is

$$\Pr(Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n) = \frac{\exp\left[\sum_{y=1}^{n} y_j(\mathbf{x}_j'\beta + \gamma)\right]}{\prod_{j=1}^{n}\left[1 + \exp(\mathbf{x}_j'\beta + \gamma)\right]}. \tag{2}$$

The usual way to make inferences about β and γ is to maximize (2) with respect to these regression coefficients.

Suppose we have interest in inferences about β, and regard γ as a nuisance parameter. Then, instead of estimating γ from the above unconditional likelihood function, we can eliminate it by conditioning on the observed value of its sufficient statistic

$$m = \sum_{j=1}^{n} y_j.$$

This yields the conditional likelihood function,

$$\Pr(Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n | m) = \frac{\exp\left(\sum_{j=1}^{n} y_j \mathbf{x}_j'\beta\right)}{\sum_R \left(\exp \sum_{j=1}^{n} y_j \mathbf{x}_j'\beta\right)}, \tag{3}$$

where the outer summation in the denominator of (3) is over the set

$$R = \left\{(y_1, y_2, \ldots, y_n): \sum_{j=1}^{n} y_j = m\right\}.$$

We can now approach inference about β in two ways: asymptotic and exact. An asymptotic approach is to maximize the conditional likelihood function (3). This is a special case of the Breslow and Day[6] method discussed in the next section for handling stratified data. Exact inference about β is based on the permutational distribution of its sufficient statistics. One can observe from the form of (3) that the $(p \times 1)$ vector of sufficient statistics for β is

$$\mathbf{t} = \sum_{j=1}^{n} y_j \mathbf{x}_j, \tag{4}$$

and its distribution is

$$\Pr(T_1 = t_1, T_2 = t_2, \ldots, T_p = t_p) = \frac{c(\mathbf{t})\, e^{\mathbf{t}'\beta}}{\sum_{\mathbf{u}} c(\mathbf{u})\, e^{\mathbf{u}'\beta}}, \tag{5}$$

where

$$c(\mathbf{t}) = |S(\mathbf{t})|,$$

$$S(\mathbf{t}) = \left\{(y_1, y_2, \ldots, y_n): \sum_{j=1}^{n} y_j = m, \sum_{j=1}^{n} y_j x_{ij} = t_i, i = 1, 2, \ldots, p\right\},$$

$|S|$ denotes the number of distinct elements in the set S, and the summation in the denominator is over all $\mathbf{u}$ for which $c(\mathbf{u}) \geqslant 1$. In other words, $c(\mathbf{t})$ is the count of the number of binary sequences of the form $(y_1, y_2, \ldots, y_n)$ such that $\sum_j y_j = m$ and $\sum_j y_j x_{ij} = t_i$ for $i = 1, 2, \ldots, p$. Exact inference about β requires computation of coefficients such as $c(\mathbf{t})$ in which some of the sufficient statistics are fixed at their observed values and others are required to vary over their permissible ranges.

2.2. Logistic regression for stratified binary data

Suppose there are N strata, with binary responses in each of them. Let the ith stratum have m_i responses and $n_i - m_i$ non-responses. For all $1 \leqslant i \leqslant N$, and $1 \leqslant j \leqslant n_i$, let $Y_{ij} = 1$ if the jth individual in the ith stratum responded; 0 otherwise. Define $\pi_{ij} = \Pr(Y_{ij} = 1 | \mathbf{x}_{ij})$ where $\mathbf{x}_{ij}$ is a p-dimensional vector of covariates for the jth individual in the ith stratum. The logistic regression model for π_{ij} is of the form

$$\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \gamma_i + \mathbf{x}'_{ij}\beta, \tag{6}$$

where γ_i is a stratum specific scalar parameter and β is a $(p \times 1)$ vector of parameters common across all N strata. Usual interest is in inferences about β, with the γ_i's regarded as nuisance parameters. One could, of course, estimate these nuisance parameters by the maximum likelihood method. The usual asymptotic theory of maximum likelihood estimation, however, requires that the dimension of the parameter space is fixed as the number of observations increase. Cox and Hinkley[8] (page 292) observe in general that when the dimension of the parameter space is large comparable to the number of observations, the MLE can have serious bias. A classic example of this situation, discussed in both Andersen[9] (page 69) and Breslow and Day[6] (page 249), is the estimation of the common odds ratio from matched pairs data. A logistic regression model for such data would contain a set of stratum specific nuisance parameters, one for each matched pair, and a single odds ratio parameter, common across all the matched pairs. If one estimates the stratum specific nuisance parameters from the data, the estimate of the common odds ratio has been shown to converge to the square of its true value.

Instead of estimating all the stratum specific parameters, an alternative approach, popularized by Breslow and Day,[6] is to eliminate these nuisance parameters by conditioning on their sufficient statistics, in this case, the number of responses, m_i, in each stratum. The conditional likelihood, or conditional probability of observing $Y_{ij} = y_{ij}, j = 1, 2, \ldots, n_i, i = 1, 2, \ldots, N$ is then

$$\Pr(Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n | m_1, m_2, \ldots, m_N) = \frac{\exp\left[\sum_{y=1}^{n} \sum_{j=1}^{n_i} y_{ij}(\mathbf{x}'_{ij}\beta)\right]}{\sum_{i=1}^{N} \sum_{R_i} \exp\left[\sum_{i=1}^{N} \sum_{j=1}^{n_i} y_{ij}(\mathbf{x}'_{ij}\beta)\right]} \tag{7}$$

where the two outer summations in the denominator are over the sets

$$R_i = \left\{(Y_{i1}, \ldots, Y_{in_i}): \sum_{j=1}^{n_i} Y_{ij} = m_i\right\},$$

for $i = 1, 2, \ldots, N$. Notice that the nuisance parameters, γ_i, have factored out of the above conditional likelihood. The Breslow and Day[6] approach is to make asymptotic inferences about β by maximizing (7). Exact inference is based on the sufficient statistics for β.

From (7) we can see that the vector of sufficient statistics for β is

$$\mathbf{t} = \sum_{i=1}^{N} \sum_{j=1}^{n_i} y_{ij}\mathbf{x}_{ij}, \tag{8}$$

and its conditional distribution is

$$\Pr(T_1 = t_1, T_2 = t_2, \ldots, T_p = t_p) = \frac{c(\mathbf{t})\, e^{\beta' \mathbf{t}}}{\sum_{\mathbf{u}} c(\mathbf{u})\, e^{\beta' \mathbf{u}}}, \tag{9}$$

where

$$c(\mathbf{t}) = |S_N(\mathbf{t})|,$$

$$S_N(\mathbf{t}) = \left\{ (y_{ij}, j = 1, \ldots, n_i, i = 1, \ldots, N): \sum_{i=1}^{N} \sum_{j=1}^{n_i} y_{ij} \mathbf{x}_{ij} = \mathbf{t}, \sum_{j=1}^{n_i} y_{ij} = m_i \right\},$$

$|S_N|$ denotes the number of distinct elements in the set S_N, and the summation in the denominator is over all $\mathbf{u}$ for which $c(\mathbf{u}) \geq 1$. In other words, $c(\mathbf{t})$ is the count of the number of ways of selecting the binary sequence $\{y_{ij}, i = 1, \ldots, N, j = 1, \ldots, n_i\}$ so as to satisfy the two conditions

$$\sum_{i=1}^{N} \sum_{j=1}^{n_i} y_{ij} \mathbf{x}_{ij} = \mathbf{t}, \tag{10}$$

and

$$\sum_{j=1}^{n_i} y_{ij} = m_i. \tag{11}$$

Notice that the distribution of T is of the same form for both stratified and unstratified logistic regression. This makes it possible to develop a single numerical algorithm for both cases.

3. EXACT CONDITIONAL INFERENCE

3.1. Conditional inference for a single parameter

Suppose without loss of generality that we wish to make inferences about the single parameter β_p. By the sufficiency principle, the conditional distribution of T_p given $t_1, t_2, \ldots, t_{p-1}$ depends only on β_p. Let $f(t_p|\beta_p)$ denote the conditional probability $\Pr(T_p = t_p | T_1 = t_1, \ldots, T_{p-1} = t_{p-1})$. Then

$$f(t_p|\beta_p) = \frac{c(t_1, t_2, \ldots, t_p)\, e^{\beta_p t_p}}{\sum_u c(t_1, t_2, \ldots, t_{p-1}, u)\, e^{\beta_p u}}, \tag{12}$$

where the summation in the denominator is over all values of u for which $c(t_1, t_2, \ldots, t_{p-1}, u) \geq 1$. Since this probability does not involve the nuisance parameters $(\beta_1, \beta_2, \ldots, \beta_{p-1})$, we can use it for inference about β_p. Notice that the above conditional probability is of the same form whether t is defined by equation (5) or by equation (9), thereby providing a unified method of exact inference for both the unstratified and stratified logistic regression models.

3.1.1. Hypothesis testing

Suppose we wish to test

$$H_0: \beta_p = 0$$

against the two-sided alternative

$$H_1: \beta_p \neq 0.$$

We obtain the exact p-value by summing (12) over some specified critical region E:

$$p = \sum_{v \in E} f(v|\beta_p = 0). \tag{13}$$

We can specify the critical region E in different ways that lead to different types of tests. Two popular tests are the 'conditional probabilities' test, and the 'conditional scores' test. In the conditional probabilities test, the critical region, denoted by E_{cp}, contains all values of the test statistic that yield a conditional probability no larger than the conditional probability at the observed value of t_p:

$$E_{cp} = \{v: f(v|\beta_p = 0) \leqslant f(t_p|\beta_p = 0)\}. \tag{14}$$

In the conditional scores test, the critical region, denoted by E_{cs}, contains all values of the test statistic whose conditional scores equal or exceed the conditional score at the observed value of the test statistic:

$$E_{cs} = \{v:(v - \mu_p)^2 \sigma_p^{-2} \geqslant (t_p - \mu_p)^2 \sigma_p^{-2}\}, \tag{15}$$

where μ_p and σ_p^2 are the mean and variance of T_p, based on its conditional distribution as specified by (12) at $\beta_p = 0$. For both types of exact tests we need an algorithm that can give us all the coefficients, $c(t_1, t_2, \ldots, t_{p-1}, v)$, with $t_1, t_2, \ldots, t_{p-1}$ fixed at their observed values, and v varying over the entire range of T_p. Once we obtain these coefficients, computation of the exact p-value is simply a matter of appropriate sorting and summing.

An asymptotic version of the conditional scores test is also possible. Here, we obtain the p-value by referring the observed score, $(t_p - \mu_p)^2 \sigma_p^{-2}$, to a chi-squared distribution on one degree of freedom. Note though that even for this asymptotic test it is necessary to compute the conditional mean, μ_p, and the conditional variance, σ_p. Asymptotic approximations to these conditional moments are available in Zelen.[10]

3.1.2. Estimation

To obtain a level-α confidence interval, (β_-, β_+) for β_p, we invert the above test. Define

$$F_1(t_p|\beta) = \sum_{v \geqslant t_p} f(v|\beta)$$

and

$$F_2(t_p|\beta) = \sum_{v \leqslant t_p} f(v|\beta)$$

Let $t_{\min}$ and $t_{\max}$ be the smallest and largest possible values of t_p in the distribution (12). The lower confidence bound, β_-, is such that

$$F_1(t_p|\beta_-) = \alpha/2 \quad \text{if } t_{\min} < t_p \leqslant t_{\max},$$

$$\beta_- = -\infty \text{ if } t_p = t_{\min}.$$

Similarly the upper confidence bound, β_+, is such that

$$F_2(t_p|\beta_+) = \alpha/2 \text{ if } t_{\min} \leqslant t_p < t_{\max},$$

$$\beta_+ = \infty \quad \text{if } t_p = t_{\max}.$$

One can show that this definition does indeed produce an interval, and the interval is guaranteed to have the desired $(100)(1-\alpha)$ per cent coverage for β_p.

We can compute a point estimate for β_p in two ways. We obtain the conditional maximum likelihood estimate, β_{cmle}, by maximizing $f(t_p|\beta)$ by choice of β. If, however, either $t_p = t_{\min}$, or if $t_p = t_{\max}$, β_{cmle} is undefined, since we cannot maximize the likelihood function. An alternative estimate for β_p that has several useful properties (see, for example, Hirji *et al.*[11]) is the median unbiased estimate

$$\beta_{\text{mue}} = (\beta_+ + \beta_-)/2,$$

where we evaluate β_- and β_+ at a confidence level $\alpha = 0{\cdot}5$. If $\beta_- = -\infty$, we define $\beta_{\text{mue}} = \beta_+$, while if $\beta_+ = \infty$, we define $\beta_{\text{mue}} = \beta_-$. Thus, unlike the maximum likelihood estimate, the median unbiased estimate is always defined, even at the extreme points of the sample space.

3.2. Conditional inference for several parameters

To make inferences about several parameters simultaneously we need the joint distribution of their sufficient statistics conditional on the observed values of the remaining sufficient statistics. Suppose we partition the $(p \times 1)$ vector of regression parameters $\boldsymbol{\beta}$ into two parts; a $(p_1 \times 1)$ component, $\boldsymbol{\beta}_1$, and a $(p_2 \times 1)$ component, $\boldsymbol{\beta}_2$. Let $\mathbf{t}_1$ and $\mathbf{t}_2$ denote the corresponding vectors of sufficient statistics. We wish to test the null hypothesis

$$H_0\colon \boldsymbol{\beta}_2 = \mathbf{0}$$

against the two-sided alternative that at least one of the elements of $\boldsymbol{\beta}_2$ is not 0. By the sufficiency principle, the conditional distribution of $\mathbf{T}_2$ given $\mathbf{T}_1 = \mathbf{t}_1$ is free of the nuisance parameters $\boldsymbol{\beta}_1$. Thus, we denote the conditional probability $\Pr(\mathbf{T}_2 = \mathbf{t}_2|\mathbf{T}_1 = \mathbf{t}_1)$ by $f(\mathbf{t}_2|\boldsymbol{\beta}_2)$, where

$$f(\mathbf{t}_2|\boldsymbol{\beta}_2) = \frac{c(\mathbf{t}_1, \mathbf{t}_2)\, e^{\boldsymbol{\beta}_2' \mathbf{t}_2}}{\sum_{\mathbf{u}} c(\mathbf{t}_1, \mathbf{u})\, e^{\boldsymbol{\beta}_2' \mathbf{u}}}, \tag{16}$$

and we take the summation in the denominator of (16) over all values of $\mathbf{u}$ for which $c(\mathbf{t}_1, \mathbf{u}) \geqslant 1$. We obtain the exact two-sided p-value for testing H_0 by summing (16) over some critical region E:

$$p = \sum_{\mathbf{v} \in E} f(\mathbf{v}|\boldsymbol{\beta}_2 = \mathbf{0}). \tag{17}$$

Again we have two types of critical regions that lead, respectively, to the conditional probabilities test and the conditional scores test. The critical region for the conditional probabilities test is

$$E_{cp} = \{\mathbf{v}: f(\mathbf{v}|\boldsymbol{\beta}_2 = \mathbf{0}) \leqslant f(\mathbf{t}_2|\boldsymbol{\beta}_2 = \mathbf{0})\}. \tag{18}$$

The critical region for the conditional scores test is

$$E_{cs} = \{\mathbf{v}: (\mathbf{v} - \mu_2)' \Sigma_2^{-1} (\mathbf{v} - \mu_2) \geqslant (\mathbf{t}_2 - \mu_2)' \Sigma_2^{-1} (\mathbf{t}_2 - \mu_2)\}, \tag{19}$$

μ_2 is the mean, and Σ_2 is the variance covariance matrix of $f(\mathbf{t}_2|\boldsymbol{\beta}_2 = \mathbf{0})$. For both types of tests we need an algorithm that can give us all the coefficients $c(\mathbf{t}_1, v)$ with $\mathbf{t}_1$ fixed and $\mathbf{v}$ varying over the entire range of $\mathbf{T}_2$.

3.3. Predictive inference

Given the unstratified logistic regression model (1), suppose we wish to compute an exact confidence interval for π_0, the probability of a response at $x = x_0$. To do so, re-write the model in the form

$$\log\left(\frac{\pi_j}{1-\pi_j}\right) = (\gamma + \mathbf{x}_0'\boldsymbol{\beta}) + (\mathbf{x}_j' - \mathbf{x}_0')\boldsymbol{\beta}. \tag{20}$$

Now, exact inference for the constant term $\gamma + \mathbf{x}_0'\boldsymbol{\beta}$ in the above re-parameterized model, based on the method described in Section 3.1, will produce an exact confidence interval for $\log(\pi_0/1-\pi_0)$. We can transform this confidence interval into a prediction interval for π_0 by applying the function $\exp(\cdot)/\{1-\exp(\cdot)\}$ to the upper and lower bounds of the confidence interval. The above procedure is easy to implement in practice. One simply shifts each $\mathbf{x}_j$ by subtracting x_0 from it. Then one estimates the constant term based on the transformed data set.

3.4. Simultaneous inference on linear combinations of parameters

For notational convenience, without loss of generality, let us represent both the unstratified and stratified logistic regression models in one common form

$$\text{logit}(\boldsymbol{\Pi}) = \mathbf{X}\boldsymbol{\beta}, \tag{21}$$

where $\text{logit}(\boldsymbol{\Pi})$ is a $n \times 1$ vector of logit response probabilities whose jth component is $\log(\pi_j/1-\pi_j)$, X is an $n \times p$ data matrix, and we have incorporated the constant terms into the $(p \times 1)$ parameter vector $\boldsymbol{\beta}$. Suppose we wish to test the hypothesis

$$H_0\colon \mathbf{C}\boldsymbol{\beta} = \mathbf{0},$$

where $\mathbf{C}$ is a $(r \times p)$ matrix of full rank. We can test H_0 by rewriting the model (21) as

$$\text{logit}(\boldsymbol{\Pi}) = \mathbf{X}\boldsymbol{\beta} + \mathbf{X}\mathbf{G}\mathbf{C}\boldsymbol{\beta}, \tag{22}$$

where $\mathbf{G}'$ is the $(r \times p)$ orthocomplement to $\mathbf{C}$, that is, $\mathbf{G}\mathbf{C} = \mathbf{0}$. After reparameterizing (22) as

$$\text{logit}(\boldsymbol{\Pi}) = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2, \tag{23}$$

where $\mathbf{X}_1 = \mathbf{X}$, $\mathbf{X}_2 = \mathbf{X}\mathbf{G}$, $\boldsymbol{\beta}_1 = \boldsymbol{\beta}$, and $\boldsymbol{\beta}_2 = \mathbf{C}\boldsymbol{\beta}$, we can test $H_0\colon \boldsymbol{\beta}_2 = \mathbf{0}$ by deriving the exact distribution of $\mathbf{T}_2|\mathbf{T}_1 = \mathbf{t}_1$ as described in Section 3.2.

4. NUMERICAL ALGORITHMS

We confine ourselves to referencing the most recent algorithmic developments for exact logistic regression, rather than to describe these algorithms in detail here. Byar and Cox[12] developed an early algorithm in which all possible binary sequences of the Y variable are enumerated exhaustively. Tritchler[2] provided a substantial improvement relative to exhaustive enumeration, using a specific application of the inverse Fourier transform algorithm of Pagano and Tritchler.[13] Tritchler's algorithm, however, only applies to models with a single covariate, with possible stratification for matched sets. Hirji *et al.*[3,4] developed a general and efficient algorithm for evaluating the permutational distribution of $\mathbf{T}_2|\mathbf{T}_1 = \mathbf{t}_1$ (see equation (16)) for unstratified data, and subsequently extended it to the stratified case. Hirji[5] has recently extended these algorithms further to allow for polytomous regression.

5. EXAMPLES

The five examples in this section illustrate various important features of exact logistic regression, and the additional insights it can provide relative to maximum likelihood inference. Example 5.1 is a good, overall introduction to exact inference. It explains why the maximum likelihood method might fail for unbalanced data sets, and how one can nevertheless obtain valid inferences by generating exact permutational distributions of sufficient statistics. This example also illustrates the computation of exact prediction intervals. Example 5.2 highlights exact simulataneous hypothesis testing of several parameters. Example 5.3 illustrates exact inference for a single parameter in a stratified setting. The analysis is equivalent to computing the exact Cochran–Armitage test of trend across several $2 \times c$ contingency tables. Example 5.4 shows how one can use the exact logistic regression framework to analyse data from a cross-over clinical trial, and how one can specialize it to encompass the exact Cochran's Q test, and extensions of McNemar's test. Finally, example 5.5 illustrates the use of exact stratified logistic regression for comparing the dose–response relationships of two drugs in a repeated measures setting. All the calculations were performed by LogXact,[14] a new statistical package for exact logistic regression.

5.1. Predictors of disease free survival for osteogenic sarcoma

In a 46-patient study of non-metastatic osteogenic sarcoma conducted by Goorin *et al.*,[15] the investigators had interest in determining the predictors for a three year disease-free interval (DFI3). The covariates of interest weere gender (SEX), any osteoid pathology (AOP), and lymphocytic infiltration (LI). The data appear in Table I.

One can show, by running Fisher's exact test on individual 2×2 contigency tables formed from cross-tabulations of DF13 with each covariate in turn, that the marginal effects of LI, SEX and AOP on DFI3 are all statistically significant at the 5 per cent level. (The respective two-sided P-values are 0·0075, 0·0259 and 0·0322.) The goal, however, is to study the effects of these three covariates simultaneously through the logistic regression model

$$\log\left(\frac{\pi_j}{1-\pi_j}\right) = \gamma + \sum_{l=1}^{3} \beta_l x_{lj}, \tag{24}$$

where, for the jth subject: $x_{1j} = 1$ if LI is present and 0 otherwise; $x_{2j} = 1$ if SEX is male and 0 otherwise; $x_{3j} = 1$ if AOP is present and 0 otherwise. Unfortunately we cannot fit the above model to the data by the method of maximum likelihood because x_{1j} is a perfect predictor. Notice that, since every subject free of lymphocytic infiltration had a three year disease-free interval, the 2×2 table of DFI3 versus LI contains a zero cell count. In this situation the log-likelihood function cannot be maximized but approaches a finite upper bound as β_1 goes to $-\infty$. Therefore we cannot evaluate the first and second derivatives of the log-likelihood at the MLE, and it is not possible to estimate β_1 or its confidence interval by conventional maximum likelihood methods.

Exact inference is possible, however, and does provide a new insight with the data. Let (t_0, t_1, t_2, t_3) denote the sufficient statistics that correspond to $(\gamma, \beta_1, \beta_2, \beta_3)$. Note that t_i is just the sum of covariate-i values over all subjects with a three year disease-free interval. Thus $t_0 = 29$, $t_1 = 19$, $t_2 = 16$ and $t_3 = 12$. The distribution of counts, $c(t_0 = 29, t_1, t_2 = 16, t_3 = 12)$, for all possible values of t_1 appears in Table II.

1. There are nearly 800 million binary sequences of the form $(y_1, y_2, \ldots, y_{46})$ implicit in Table II. Yet, there are only eight distinct vectors of sufficient statistics of the form

Table I. Osteogenic sarcoma data

LI	SEX	AOP	Proportion DFI3
0	0	0	3/3 (100%)
0	0	1	2/2 (100%)
0	1	0	4/4 (100%)
0	1	1	1/1 (100%)
1	0	0	5/5 (100%)
1	0	1	3/5 (60%)
1	1	0	5/9 (56%)
1	1	1	6/17 (35%)

Table II. Exact conditional distribution for osteogenic sarcoma data

t_1	$c(29, t_1, 16, 12)$
19	29,445,360
20	147,312,480
21	271,271,448
22	231,819,344
23	95,325,644
24	17,473,144
25	1,204,008
26	19,448
Total	793,870,896

($t_0 = 29, t_1, t_2 = 16, t_3 = 12$), because so many different binary sequences yield the same values for the sufficient statistics. In operations research terminology this phenomenon is known as 'clubbing' (perhaps because different binary sequences belong to the same club, or vector of sufficient statistics). A good algorithm must exploit this clubbing, because otherwise, exhaustive enumeration of all possible binary sequences is computationally explosive.

2. The exact conditional distribution of T_1 is extremely asymmetric. Normal approximations would not work too well, though Edgeworth and saddlepoint approximations might be worth trying.
3. The observed value, $t_1 = 19$, is at the minimum of its range. This is the reason for the failure of the maximum likelihood method to produce estimates of the regression parameters.

Exact inference about the parameter that corresponds to LI is now straightforward. The exact one-sided p-value for testing $\beta_1 = 0$ is

$$p = 29445360/793870896 = 0{\cdot}037.$$

Table III. Parameter estimates for osteogenic sarcoma data

Parameter	Point estimate	Exact 95% CI	Exact P-value
γ	3·535	1·477 to ∞	0·0001
γ^*	− 0·737	− 1·910 to 0·310	0·164
β_1	− 1·886	$-\infty$ to 0·160	0·061
β_2	− 1·548	− 4·025 to 0·363	0·117
β_3	− 1·156	− 2·997 to 0·512	0·154

Since t_1 is at its minimum value, the lower 95 per cent confidence bound for β_1 is $-\infty$. The upper 95 per cent confidence bound, β_+, is the solution to

$$\frac{c(29,19,16,12)\,e^{19\beta_+}}{\sum_{t_1=19}^{26} c(29,t_1,16,12)\,e^{t_1\beta_+}} = 0{\cdot}025\,.$$

Binary search rapidly yields $\beta_+ = 0{\cdot}16$.

The conditional distributions of the other sufficient statistics and the corresponding parameter estimates obtain similarly. We can also translate the data for predictive inference, as discussed in Section 3.3, and then re-compute the parameter estimates. The results are shown in Table III. The p-values are all two-sided and based on the exact conditional scores test defined by equations (13) and (15). The parameter γ^* is the constant term of the model (24) after translating the data by subtracting 1 from x_{lj} for all l and j. Unlike the Fisher exact tests conducted on each variable separately, these results reveal that in a regression analysis, taking into account all three variables simultaneously, LI is marginally significant at the 0·06 level, while SEX and AOP are not. The exact confidence intervals for γ and γ^* have a particularly useful interpretation. Observe that the lower confidence bound for γ is 1·48. Following the discussion in Section 3.3, the probability that the most favourable subjects (females with no lymphocytic infiltration or ostoid pathology) will have a three year disease-free interval is at least

$$\frac{\exp(1{\cdot}48)}{1+\exp(1{\cdot}48)} = 0{\cdot}814\,,$$

with 95 per cent confidence. Similarly, by exponentiating the confidence interval for γ^*, we can guarantee with 95 per cent confidence that the probability of a three year disease-free interval for the most unfavourable subjects (males with lymphocytic infiltration and ostoid pathology) is between 0·128 and 0·576.

5.2. Advance indicators of HIV infection in infants

We are grateful to Dr. Shengan Lai, University of Miami, for providing this example. A hospital based prospective study of perinatal infection and human immunodeficiency virus (HIV-1) by Hutto *et al.*[16] investigated, among other things, the possibility that the CD4 and CD8 blood serum levels measured in infants at 6 months of age might predict their eventual development of HIV infection. The data on HIV infection rates and blood serum levels are displayed in Table IV.

We wish to determine through logistic regression if the CD4 nd CD8 serum levels predict HIV positivity. Now, although we have coded each covariate at three ordered levels (0, 1, 2), the

Table IV. Data on advance indicator of HIV

CD4	CD8	Proportion HIV
0	2	1/1 (100%)
1	2	2/2 (100%)
0	0	4/7 (57%)
1	1	4/12 (33%)
2	2	1/3 (33%)
1	0	2/7 (29%)
2	0	0/2 (0%)
2	1	0/13 (0%)

investigators preferred to include them in the regression model as qualitative or 'factor' variables rather than as quantitative variables since the actual numerical values of the CD4 and CD8 counts were unavailable. This requires that we split each of CD4 and CD8 into two dummy variables (0 versus 2, and 1 versus 2) in the regression model. We can specify the model formally as:

$$\log\left(\frac{\pi_j}{1-\pi_j}\right) = \gamma + \sum_{l=1}^{4} \beta_l x_{lj}, \tag{25}$$

where, for the jth subject: $x_{1j} = 1$ if CD4 is at level 0 and 0 otherwise; $x_{2j} = 1$ if CD4 is at level 1 and 0 otherwise; $x_{3j} = 1$ if CD8 is at level 0 and 0 otherwise; $x_{4j} = 1$ if CD8 is at level 1 and 0 otherwise.

As is often the case for data sets with small sample sizes or unbalanced structure, we cannot estimate the regression parameters in model (25) by the maximum likelihood method because the observed data fall on the boundary of the sample space; one will discover that conventional software packages are unable to produce any logistic regression output for this model. Nevertheless, the observed rates of HIV infection do vary considerably with the serum levels and formal tests of significance would be useful. The exact conditional distributions of appropriate sufficient statistics enable us to perform such tests. To determine if the CD8 levels predict HIV infection, we must test the null hypothesis

$$H_0\colon \beta_3 = \beta_4 = 0.$$

The sufficient statistic for β_l is $T_l = \sum x_{lj} Y_j$, and the sufficient statistic for the constant term is $T_0 = \sum Y_j$, the summation taken over all subjects. An exact test of H_0 is based on $f(t_3, t_4 | \beta_3 = \beta_4 = 0)$, the null permutational distribution of (T_3, T_4) given that the remaining sufficient statistics are fixed at their observed values, that is, $(T_0 = 14, T_1 = 5, T_2 = 8)$.

For testing H_0, we use the exact conditional scores test. For each (t_3, t_4) in the sample space of the conditional distribution, one can compute a conditional score of the form

$$q = ((t_3, t_4) - (\mu_3, \mu_4)) \sum_{3,4}^{-1} ((t_3, t_4) - (\mu_3, \mu_4))'$$

where μ_3 is the mean of T_3, μ_4 is the mean of T_4 and $\sum_{3,4}$ is the variance–covariance matirx of $f(t_3, t_4 | \beta_3 = \beta_4 = 0)$. The observed value of (t_3, t_4) is $(6, 4)$, and hence the observed conditional score is $q = 7{\cdot}293$. The critical region for the conditional scores test, E_{cs} (defined by equation (19)),

Table V. *P*-values for HIV data

Conditional score test	CD4	CD8
Exact *P*-value	0·007	0·0256
Asymptotic *P*-value	0·009	0·0261

thus consists of all (t_3, t_4) points in the sample space with conditional scores greater than or equal to 7·293. The exact p-value is

$$p_{cs} = \sum_{E_{cs}} f(t_3, t_4 | \beta_3 = \beta_4 = 0) = 0{\cdot}0256,$$

while the asymptotic p-value is obtained as the tail area to the right of 7·293 from a chi-square distribution with 2 degrees of freedom. In Table V we display the exact and asymptotic p-values for CD4 and CD8, based on the conditional scores tests. Despite the small sample size the exact and asymptotic results are very similar. The accuracy of the asymptotic results are attributable to the conditional rather than the unconditional scores statistic being referred to the chi-square distribution. This example also demonstrates that it is possible to perform asymptotic hypothesis tests on model parameters, using the scores test, even when the full model cannot be fit by the maximum likelihood method.

5.3. Schizophrenia and birth complications

We thank Dr. Armando Garsd for providing this example. A case-control study (Garsd[17]) was sought to determine the role of birth complications in people with schizophrenia. The sample consisted of 7 families with several siblings per family. An individual within a family was classified either as normal or a person with schizophrenia. There was a 'birth-complications index' available for each individual, ranging in value from 0 (uncomplicated birth) to 15 (severely complicated birth). The data are shown in Table VI. As a point of clarification we note that there are no multiple births depicted in this table. For example, the three births listed in Family 1, all with birth complications indices of 2, and all free of schizophrenia, represent three single births at three different time points.

Is there a positive correlation between the chance of schizophrenia and the birth-complications index? The data do indeed suggest some such tendency, but, the numbers are small, and the magnitude of the effect appears to vary across families. This is an ideal situation for exact logistic regression on matched sets. By treating each family as a separate matched set, one can model π_{ij}, the probability of schizophrenia for the jth sibling in the ith family in terms of the birth-complications index, x_{ij}:

$$\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \gamma_i + \beta \mathbf{x}_{ij}. \tag{26}$$

We eliminate the nuisance parameters γ_i, corresponding to the family effect, by conditioning on the total number of schizophrenics within each family. We then estimate β by the methods of Section 3.1. The results are displayed in Table VII.

The exact *P*-value is based on the conditional scores test defined by equations (13) and (15). The three asymptotic *P*-values in the above table have been derived from the likelihood function that

Table VI. Data on schizophrenia and patient complications

Family	BC index	Proportion with schizophrenia
1	15	1/1 (100%)
1	7	0/1 (0%)
1	6	0/1 (0%)
1	5	0/1 (0%)
1	3	0/2 (0%)
1	2	0/3 (0%)
1	0	0/1 (0%)
2	2	1/1 (100%)
2	0	0/1 (0%)
3	9	1/1 (100%)
3	2	0/1 (0%)
3	1	0/1 (0%)
4	2	1/1 (100%)
4	0	0/4 (0%)
5	6	0/1 (100%)
5	3	1/1 (100%)
5	0	1/1 (100%)
6	3	0/1 (0%)
6	0	1/4 (25%)
7	6	1/1 (100%)
7	2	0/1 (0%)

Table VII. Inference about the beta coefficient for schizophrenia data

Conditional maximum likelihood estimate	0·325
Exact 95 per cent confidence interval	(0·0223 to 0·741)
Asymptotic 95 per cent confidence interval	(−0·004 to 0·654)
Exact *P*-value (conditional scores)	0·0167
Asymptotic *P*-value (scores)	0·0129
Asymptotic *P*-value (Wald)	0·0528
Asymptotic *P*-value (likelihood ratio)	0·023

corresponds to model (26) in the usual manner for scores, Wald and likelihood ratio tests, respectively (see, for example, McCullagh and Nelder[18]), with one difference, they derive from the conditional likelihood function (7) rather than the unconditional likelihood function (2). This makes the asymptotic results more comparable to the exact ones since we do not lose degrees of freedom in estimating stratum specific nuisance parameters. The appropriate likelihood equations for these conditonal asymptotic tests appear in Appendix A of the LogXact[14] manual. For this small data set there are noticeable *P*-value differences between the three asymptotic tests and one must rely on the exact conditional scores test to furnish a 'gold-standard' *P*-value. Note also that this exact test is equivalent to the Cochran–Armitage exact test of trend on stratified contingency tables (see, for example, Breslow and Day,[6] Section 4.5) where one regards the data on each family as a separate stratum.

Table VIII. Cross-over data on analgesic efficacy

Patient	Drug sequence	Response P1	P2	P3
1	ABC	0	1	1
7	ABC	0	1	1
2	BCA	0	1	1
8	BCA	0	0	0
3	CAB	1	0	0
9	CAB	1	0	1
4	CBA	1	0	1
10	CBA	1	0	0
5	ACB	0	0	0
11	ACB	0	1	0
6	BAC	1	0	0
12	BAC	0	0	1

5.4. Cross-over clinical trial of analgesic efficacy

The data in Table VIII come from a three-treatment three-period cross-over clinical trial. The three drugs are A = new drug, B = aspirin, C-placebo. The primary endpoint was analgesic efficacy, here dichotomized as 0 for relief and 1 for no-relief. See Snapinn and Small[19] for details.

The question is whether the three treatments differ. We answer this question by including treatment as the primary covariate in a logistic regression model for matched sets. In this model, we include treatment as an unordered categorical covariate at three levels, and hence, with two degrees of freedom. We regard each patient as a matched set. Within such a matched set there are three observed responses, one at each of the three time periods P1, P2 and P3. Now although these responses are all on the same patient, and are therefore dependent, we assume that we can remove this dependence by appropriate modelling as in Jones and Kenward.[20] For the present data set we assume that we can regard the three response probabilities within a matched set as independent if they arise in a logistic regression model that contains a stratum specific constant and covariate terms for treatment and period effects. Since there are three periods, we capture the period effect with two degrees of freedom. Technically the model should also include a two degree of freedom covariate term for the carry-over effect. For this small data set, however, the period effect and the carry-over effect are aliased, that is, there are insufficient data points to distinguish between the parameters that correspond to period from those that correspond to carry-over. We may thus specify the model as:

$$\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \gamma_i + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \beta_3 x_{3ij} + \beta_4 x_{4ij}, \tag{27}$$

where: γ_i is the stratum effect for the ith matched set (or subject); x_{1ij} is a dummy variable that assumes a value 1 if drug A was administered to subject i in period j, 0 otherwise; x_{2ij} is a dummy variable that assumes a value 1 if drug B was administered to subject i in period j, 0 otherwise; x_{3ij} is a dummy variable that assumes a value 1 if period j is P1, 0 otherwise; x_{4ij} is

Table IX. Analysis of analgesic efficacy data

Type of test	Chi-squared value	*P*-value
Likelihood ratio (asymptotic)	8·7378	0·0127
Wald (asymptotic)	5·0875	0·0786
Scores (asymptotic)	7·8010	0·0202
Conditional scores (exact)	7·0634	0·0289

a dummy variable that assumes a value 1 if period j is P2, 0 otherwise. The results for the two degree of freedom test

$$H_0: \beta_1 = \beta_2 = 0$$

that there is no treatment effect are shown in Table IX.

As was the case with example 5.3, the above asymptotic tests all derive from conditional likelihood function (7) for model (27) rather than the unconditional likelihood function (2). The conditional likelihood function is free of the stratum specific nuisance parameters, γ_i. There is a fairly large discrepancy between the Wald and likelihood ratio tests so that one would prefer to rely on the exact conditional scores test.

We next computed the two degree of freedom test

$$H_0: \beta_3 = \beta_4 = 0$$

that there is no period effect. The exact conditional scores *P*-value for this test was 0·8842, implying that we could drop the two terms that correspond to period from the model (27). Accordingly, we did drop these two terms and once again tested the hypothesis

$$H_0: \beta_1 = \beta_2 = 0$$

this time from the reduced model

$$\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \gamma_i + \beta_1 x_{1ij} + \beta_2 x_{2ij}. \tag{28}$$

This time the exact conditional scores *P*-value was 0·026, not very different from the *P*-value when we included period in the model, thus confirming that the period effect is not significant. Note that the above two-degree of freedom exact test for a drug effect in model (28) is the exact analogue of Cochran's Q test (for example, see Siegel and Castellan,[21] page 170).

Finally, we computed separate exact tests of $\beta_1 = 0$ (*P*-value equals 0·0159) and $\beta_2 = 0$ (*P*-value equals 0·0972) from model (28). We can regard these two tests as exact extensions of McNemar's test, since they deal with the comparison of two repeated measures on each subject, while adjusting for a third repeated measure through regression.

5.5. Bupenorphine treatment for drug addicts

We thank Dr. Edward Lee, Substance Abuse Treatment Unit, Department of Psychiatry, Yale University, for providing this example of multiple binary response on five substance abusers. The five individuals were treated with both the control drug ($X = 1$) and Bupenorphine ($X = 2$) at each of four doses (0, 0·125, 0·250, 0·500 mg/m^2). The binary response measured at each does level was presence/absence of abnormal heartbeat ($Y = 1/0$). The data are displayed in Table X.

Table X. Data on substance abusers and abnormal heartbeat

Patient	Dose	Treatment 1 response	Treatment 2 response
1	0	0	1
2	0	0	0
3	0	0	1
4	0	0	0
5	0	0	0
1	125	0	1
2	125	1	1
3	125	0	1
4	125	0	1
5	125	1	1
1	250	1	1
2	250	1	1
3	250	1	1
4	250	1	1
5	250	1	1
1	500	1	1
2	500	1	1
3	500	1	1
4	500	1	1
5	500	1	1

The question of interest was whether Bupenorphine increased the probability of abnormal heartbeat relative to the control drug. The data were complicated by the fact that each individual was treated several times, at different dose levels of both treatments, thereby providing a sequence of clustered binary responses. We handled this problem by using stratified logistic regression, regarding each individual as a separate stratum. The model was thus

$$\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \gamma_i + \beta_1 x_{1ij} + \beta_2 x_{2ij}. \tag{29}$$

where i indexes the strata, $i = 1, 2 \ldots, 5$, j indexes the different dose levels within each stratum, $j = 1, 2, \ldots, 4$, x_{1ij} is the jth drug dose in the ith stratum, and x_{2ij} is the jth treatment in the ith stratum (1 for control drug; 2 for Bupenorphine).

As with several other examples presented in this paper, the maximum likelihood method failed to produce estimates of the regression coefficients. The exact method, however, based on the permutation distribution of the sufficient statistics produced parameter estimates as shown in Table XI.

The exact method reveals that Bupenorphine does indeed induce a statistically significant increase in abnormal heartbeat, after adjusting for the effects of clustering, and varying dose levels. The coefficient β_1 is a trend parameter that measures the amount by which the odds of abnormal heartbeat increase for a unit increase in the dose. For instance, an increase in the dose by 0·125 mg/m^2 increases the odds of abnormal heartbeat by a factor $\exp(0{\cdot}021 \times 125) = 13{\cdot}8$. The coefficient β_2 reveals that the odds of abnormal heartbeat increase by a factor $\exp(2{\cdot}22) = 9{\cdot}2$, if the patient switches from the control drug to Bupenorphine. This

Table XI. Analysis of substance abusers data

Parameter	Point estimate	Exact 95% CI	Exact P-value
β_1	0·0210	0·00825 to ∞	$5{\cdot}11 \times 10^{-6}$
β_2	2·22	0·0977 to ∞	0·0396

interpretation is troublesome at dose 0, however. As one of the referees has noted, switching from the control drug to the placebo at dose 0 actually amounts to being treated by placebo in either case, and should have no effect on the heartbeat. Perhaps an interaction term should be included in the model, but there are too few differences in response between the two treatments to support such a model.

6. CONCLUSIONS

We have provided a way to analyse small-sample binary data with covariates, and have illustrated our approach through several examples that one could not analyse accurately by conventional methods of logistic regression. For data in the form of independent binary observations, we used the unstratified logistic regression model, and based our inference on appropriate permutational distributions of the sufficient statistics. For clustered binary data, consisting of a few experimental units, and repeated binary observations on each unit, we used the stratified logistic regression model. We treated each experimental unit as a separate stratum or matched set. The inference on the regression parameters proceeded as before and was based on permutational distributions of sufficient statistics. The permutational approach for clustered binary data is a useful complement to the generalized estimating equations (GEE) approach of Zeger and Liang,[22] for it is valid in small samples while the latter is valid in large samples. We note, however, that the permutational approach is conditional whereas the GEE approach is marginal, leading to slightly different interpretations for the parameter estimates.

We have seen that for data sets with small sample sizes or unbalanced structure the conventional maximum likelihood approach may fail, even though the covariates in the model are statistically significant. The permutational approach on the other hand provides valid inferences for this situation. Example 5.1 was an instance of failure of the maximum likelihood approch because of a zero cell count in a 2×2 table formed by the response variable and a binary covariate. Actually, the maximum likelihood method can fail under even weaker conditions. There is a fine discussion of these conditions in Santner and Duffy,[23] page 234, with many related references.

In sparse data settings where the maximum likelihood estimates do exist, one can compare the exact and asymptotic results. Sometimes the two results are similar and at other times they differ considerably. It would be useful to provide a simple rule of thumb (analogous to Cochran's conditions) for identifying when the exact and asymptotic results are likely to be similar. This research remains to be done however. Another fruitful area for further research is attainment of accurate asymptotic approximations for conditional distributions of the form (16), their tail areas, and their moments. Recent work on saddlepoint, Edgeworth and Gibbs–Skovgaard approximations by Pierce and Peters[24] and Kolassa and Tanner,[25] and on Markov chain Monte Carlo sampling by Geyer and Thompson,[26] Diaconis and Sturmfels,[27] and Foster, Mc Donald and Smith[28] show considerable promise in this regard.

ACKNOWLEDGEMENTS

The authors thank Byron Jones, Sander Greenland and two referees for their careful reading of the manuscript and their constructive suggestions. We are grateful to Gary Koch and Juha Alho who independently suggested transforming the regression model and thereby performing exact tests on linear combinations of parameters.

REFERENCES

1. Cox, D. R. and Snell, E. J. *Analysis of Binary Data*, 2nd edn, Chapman and Hall, London, 1989.
2. Tritchler, D. 'An algorithm for exact logistic regression, *Journal of the American Statistical Association*, **79**, 709–711 (1984).
3. Hirji, K. E., Mehta, C. R. and Patel, N. R. 'Computing distributions for exact logistic regression', *Journal of the American Statistical Association*, **82**, 1110–1117 (1987).
4. Hirji, K. F. Mehta, C. R. and Patel, N. R. 'Exact inference for matched case-control studies', *Biometrics*, **44**, 803–814 (1988).
5. Hirji, K. F. 'Exact distributions for polytomous data', *Journal of the American Statistical Association*, **87**, 487–492 (1992).
6. Breslow, N. E. and Day, N. E. *Stat Methods in Cancer Research*, IARC, Lyon, 1980.
7. Gail, M. H., Lubin, J. H. and Rubenstein, L. V. 'Likelihood calculations for matched case-control studies and survival studies with tied death times', *Biometrika*, **68**, 703–707 (1981).
8. Cox, D. R. and Hinkley, D. V. *Theoretical Statistics*, Chapman and Hall, London, 1974.
9. Andersen, E. B. *Conditional Inference and Models for Measuring*, Mental Hygienisk Forlag, Copenhagen, 1973.
10. Zelen, M. 'Multinomial response models', *Computational Statistics and Data Analysis*, **12**, 249–254.
11. Hirji, K. F., Tsiatis, A. A. and Mehta, C. R. 'Median unbiased estimation for binary data', *The American Statistician*, **43**, 7–11 (1989).
12. Byar, L. and Cox, C. 'Algorithm AS142', *Applied Statistics*, **28**, 319–324 (1979).
13. Pagano, M. and Tritchler, D. 'Permutation distributions in polynomial time', *Journal of the American Statistical Association*, **78**, 435–440 (1983).
14. LogXact. *Software for Exact Logistic Regression*, Cytel Software Corporation, Cambridge, MA, 1992.
15. Goorin, A. M., Perez-Atayde, A., Gebhardt, M. and Andersen, J. 'Weekly high-dose methotrexate and doxorubicin for osteosarcoma, *Journal of Clinical Oncology*, **5**, 1178–1184 (1987).
16. Hutto, C., Parks, W. P. and Lai, S. 'A hospital based prospective study of perinatal infection with HIV-1', *Journal of Pediatrics*, **118**, 347–353 (1991).
17. Garsd, A. 'Schizophrenia and birth complications', unpublished manuscript, 1988.
18. McCullagh, P. and Nelder, J. A. *Generalized Linear Models*, 2nd edn, Chapman and Hall, London, 1989.
19. Snapinn, S. M. and Small, R. D. 'Regression models for categorical data', *Biometrics*, **42**, 583–592 (1986).
20. Jones, B. and Kenward, M. G. 'Binary data from a three-period trial', *Statistics in Medicine*, **6**, 555–564 (1987).
21. Seigel, S. and Castellan, N. J. *Nonparametric Statistics for the Behavioral Sciences*, 2nd edn, New York, 1988.
22. Zeger, S. L. and Liang, K. Y. 'Longitudinal analysis', *Biometrics*, **42**, 121–130 (1986).
23. Santner, T. J. and Duffy, D. E. *The Statistical Analysis of Discrete Data*, Springer-Verlag, New York, 1989.
24. Pierce, D. A. and Peters, D. 'Practical use of higher order asymptotics for multiparameter exponential families (with discussion)', *Journal of the Royal Statistical Society, Series B*, **54**, 3, 701–737 (1992).
25. Kolassa, J. E. and Tanner M. A. 'Approximate conditional inference in exponential families via the Gibbs sampler', *Journal of the American Statistical Association*, **89**, 426, 697–702 (1994).
26. Geyer, C. J. and Thompson, E. A. 'Constrained Monte Carlo maximum likelihood for dependent data', *Journal of the Royal Statistical Society, Series B*, **54**, (3), 657–699 (1992).
27. Diaconis, P. and Sturmfels, B. 'Algebraic algorithms for sampling from conditional distributions', Technical Report, Department of Mathematics, Harvard University, 1993.
28. Foster, J. J., Mc Donald, J. W. and Smith, P. W. F. 'Monte Carlo exact conditional tests for log-linear and logistic models', Technical Report, Department of Social Statistics, University of Southampton, Southampton, U.K., 1995.