


11-1-2012

# Exact Logistic Regression for a Matched Pairs case-Control Design with Polytomous Exposure Variables

Shyam S. Ganguly

*Sultan Qaboos University, Muscat 123, Oman*

Follow this and additional works at: <http://digitalcommons.wayne.edu/jmasm>

 Part of the [Applied Statistics Commons](#), [Social and Behavioral Sciences Commons](#), and the [Statistical Theory Commons](#)

## Recommended Citation

Ganguly, Shyam S. (2012) "Exact Logistic Regression for a Matched Pairs case-Control Design with Polytomous Exposure Variables," *Journal of Modern Applied Statistical Methods*: Vol. 11: Iss. 2, Article 16.

Available at: <http://digitalcommons.wayne.edu/jmasm/vol11/iss2/16>

This Regular Article is brought to you for free and open access by the Open Access Journals at DigitalCommons@WayneState. It has been accepted for inclusion in Journal of Modern Applied Statistical Methods by an authorized administrator of DigitalCommons@WayneState.

## Exact Logistic Regression for a Matched Pairs Case-Control Design with Polytomous Exposure Variables

Shyam S. Ganguly  
Sultan Qaboos University  
Muscat 123, Oman

---

Logistic regression methods are useful in estimating odds ratios under matched pairs case-control designs when the exposure variable of interest is binary or polytomous in nature. Analysis is typically performed using large sample approximation techniques. When conducting the analysis with polytomous exposure variable, situations where the numbers of discordant pairs in the resulting cells are small or the data structure is sparse can be encountered. In such situations, the asymptotic method of analysis is questionable, thus an exact method of analysis may be more suitable. A method is presented that performs exact inference in the case of pair-wise matched case-control data with more than two unordered exposure categories using a distribution of conditional sufficient statistics of logistic model parameters.

**Key words:** Conditional logistic regression, sufficient statistic, exact analysis, Diophantine systems.

---

### Introduction

In epidemiological studies, the matched case-control design is often conducted to establish the relationship between disease incidence and an exposure variable of interest in terms of odds ratio (Mantel & Haenszel, 1959; Miettinen, 1970; Ejigou & McHugh, 1977, 1981). The binary logistic model (Cox, 1970) is useful in the estimation of odds ratios under matched pair case-control designs. Prentice (1976), Holford (1978), Holford, et al. (1978), Klinbaum, et al. (1982) and Breslow and Day (1980) provide detailed discussions regarding the estimation of odds ratios using binary logistic models that are conditional on disease status. The polytomous logistic model (Prentice & Pyke, 1979; Dubin & Pasternack, 1988; Liang & Stewart, 1987) has also been found to be useful in estimating odds ratios in the case of matching design when multiple case-control groups are considered; however, when conducting a pair-wise matched

case-control study, a situation where the risk factor under investigation has more than two levels, which may be ordinal or nominal in nature, can be encountered. Ganguly and Naik-Nimbalkar (1995) discuss analysis in the case of a risk factor with a natural ordering, and Ganguly (2006) further estimated the covariate adjusted odds ratios in the case of the ordinal multiple level exposure variables.

Nominal response situations were studied in detail by Pike, et al. (1975), who estimated odds ratios between blood types and development of disease, considering a hypothetical data set. Holford, et al. (1978) analyzed the same data set using a binary logistic model with a conditional likelihood procedure, and Ganguly and Naik-Nimbalkar (1992a, 1992b) further analyzed the data, modeling retrospective probabilities using a polytomous logistic model. All estimation procedures described are based on maximizing the conditional likelihood that relies on asymptotic approximations. The validity of the analysis based on the asymptotic method may be in question when the sample size is small or the data are sparse. In such situations the exact method of analysis is more appropriate (Breslow & Day, 1980; Agresti, 1990; Mehta, 1994).

---

Shyam S. Ganguly is an Associate Professor of Statistics & Epidemiology in the College of Medicine and Health Sciences in the Department of Family Medicine & Public Health. Email him at: [ganguly@squ.edu.om](mailto:ganguly@squ.edu.om).

Cox (1970) put forth a method for exact logistic regression analysis involving a single parameter in unmatched logistic models, which can also be applied to matched designs when the response is binary. Tritchler (1984) estimated model parameters based on an algorithm developed for a permutation test by Pagano and Tritchler (1983). Hirji, Mehta and Patel (1988) developed a recursive algorithm to compute the exact conditional distribution of sufficient statistics of the parameters involved in a logistic model for analyzing data from a matched case-control design. Hirji (1992) provided an efficient method for computing exact conditional distributions of sufficient statistics for the parameters involved in polytomous response models. However, none of these studies discuss matched case-control designs involving more than two exposure categories. This article proposes a method that uses the conditional distribution of sufficient statistics of logistic model parameters to perform exact inference in the case of a 1-1 matched case-control data with a polytomous exposure variable.

The Logistic Regression Model

Assume  $k$  possible levels of an exposure variable of interest. Let  $p_{ji}$  be the probability that, in a given pair, the case is exposed to level  $j$  and the control is exposed to level  $i$ , conditional on one of them being exposed to level  $j$  and the other exposed to level  $i$  ( $1 \leq i < j \leq k$ ). In addition, let  $F_1$ , be the exposure level associated with the case and  $F_0$  for the control. Consider  $n_{ij}$  as the number of case-control pairs in which the case is exposed to level  $i$  and the control is exposed to level  $j$  and assume that the exposure levels associated with the case and the control are independent. The results of the case-control investigation, in general, may be represented as shown in Table 1.

Let  $s_{ij} = n_{ij} + n_{ji}$  ( $1 \leq i < j \leq k$ ) represent the number of discordant pairs referencing the  $(i, j)^{th}$  and  $(j, i)^{th}$  cells in Table 1. Further, consider  $Y_{ij\ell}$  as the case-control indicator for the  $\ell^{th}$  pair ( $\ell = 1, \dots, s_{ij}$ ) such that

$$Y_{ij\ell} = \begin{cases} 1 & \text{when case is at } i^{th} \text{ and control is at } j^{th} \text{ level} \\ 0 & \text{otherwise} \end{cases}$$

Table 1: Representation of Data from a Matched Pair Study with  $k$  Exposure Levels

Exposure Level for Case ( $F_1$ )	Exposure Level for Control ( $F_0$ )				
	1	2	$i$	$j$	$k$
1	$n_{11}$	$n_{12}$	$n_{1i}$	$n_{1j}$	$n_{1k}$
2	$n_{21}$	$n_{22}$	$n_{2i}$	$n_{2j}$	$n_{2k}$
$i$	$n_{i1}$	$n_{i2}$	$n_{ii}$	$n_{ij}$	$n_{ik}$
$j$	$n_{j1}$	$n_{j2}$	$n_{ji}$	$n_{jj}$	$n_{jk}$
$k$	$n_{k1}$	$n_{k2}$	$n_{ki}$	$n_{kj}$	$n_{kk}$

Following Ganguly and Naik-Nimbalkar (1992a) the probability that, in a given pair, the case is exposed to level  $j$  and the control is exposed to level  $i$  conditional on the fact that one is exposed to level  $j$  and the other is exposed to level  $i$ , is given by

$$p_{jil} = p_{ji} = \frac{\exp(\alpha_j - \alpha_i)}{1 + \exp(\alpha_j - \alpha_i)}, \quad (1)$$

with

$$P_{ji+} + p_{ij} = 1, 1 \leq i < j \leq k, \quad \ell = 1, \dots, s_{ij}.$$

The parameter  $\alpha_j$  ( $j = 1, \dots, k$ ) describes the additional exposure for an individual in the  $j^{th}$  category for becoming a case. The odds ratios, for comparing categories  $j$  and  $i$ , under model (1) is given by

$$r_{ji} = \exp(\alpha_j - \alpha_i), (1 \leq i < j \leq k).$$

Exact Conditional Distribution of Sufficient Statistics

If the observed discordant case-control pairs in the  $(i, j)^{th}$  and  $(j, i)^{th}$  are considered as

$y_{ij1}, y_{ij2}, \dots, y_{ij\ell} \dots y_{ijs_{ij}}$  and assumed to be independent, then the likelihood  $L_{ij}$  for the  $s_{ij}$  pairs is conditional on the study design and is given by

$$L_{ij} = \text{pr} \left[ Y_{ij1} = y_{ij1}, \dots, Y_{ij\ell} = y_{ij\ell}, \dots, Y_{ijs_{ij}} = y_{ijs_{ij}} \mid S_{ij} = s_{ij} \right]$$

$$= \prod_{\ell=1}^{s_{ij}} (p_{ij\ell})^{y_{ij\ell}} (1 - p_{ij\ell})^{1 - y_{ij\ell}} \tag{2}$$

Using relation (1) and (2),  $L_{ij}$  is given by

$$L_{ij} = \frac{\exp \left[ n_{ij} (\alpha_i - \alpha_j) \right]}{\prod_{\ell=1}^{s_{ij}} \left[ 1 + \exp(\alpha_i - \alpha_j) \right]}, \tag{3}$$

therefore, the overall conditional likelihood  $L$  is given by

$$L = \prod_{i=1}^{k-1} \prod_{j=i+1}^k L_{ij}$$

$$= \frac{\exp \left[ \sum_{i=1}^{k-1} \alpha_i t_i \right]}{\prod_{i=1}^{k-1} \prod_{j=i+1}^k \left[ 1 + \exp(\alpha_i - \alpha_j) \right]^{s_{ij}}}, \tag{4}$$

where

$$\sum_{i=1}^k \alpha_i = 0$$

or

$$\alpha_k = - \sum_{i=1}^{k-1} \alpha_i$$

and

$$t_i = \sum_{j=i+1}^k n_{ij} - \sum_{j=1}^{i-1} n_{ji} + \sum_{j=1}^{k-1} n_{jk},$$

$$i = 1, \dots, k - 1.$$

The likelihood (4) can also be represented as

$$L = H(\alpha_1, \dots, \alpha_{k-1}) \left\{ \exp \left( \sum_{i=1}^{k-1} \alpha_i t_i \right) \right\}, \tag{5}$$

where

$$H(\alpha_1, \dots, \alpha_{k-1}) = \frac{1}{\prod_{i < j} \left[ 1 + \exp(\alpha_i - \alpha_j) \right]^{s_{ij}}}$$

Relation (5) shows an exponential family of dimension  $k-1$  and the  $T_i$ 's are jointly sufficient for  $\alpha_i$  ( $i = 1, \dots, k-1$ ), whose joint distribution is obtained by summing over all  $n_{ij}$  values, such that  $T_i = t_i$  ( $i = 1, \dots, k-1$ ) and  $S_{ij} = s_{ij}$ . The joint distribution is thus given by

$$\text{pr} (T_1 = t_1, \dots, T_{k-1} = t_{k-1} \mid S_{ij} = s_{ij}, i < j = 1, \dots, k)$$

$$= \frac{C(t_1, \dots, t_{k-1}) \exp \left[ \sum_{i=1}^{k-1} \alpha_i t_i \right]}{\prod_{i=1}^{k-1} \prod_{j=i+1}^k \left[ 1 + \exp(\alpha_i - \alpha_j) \right]^{s_{ij}}}, \tag{6}$$

where  $C(t_1, \dots, t_{k-1})$  is the number of distinct set of values assumed by  $n_{ij}$  which yield the values  $t_1, \dots, t_{k-1}$  for the joint sufficient statistic.

Following Cox (1970), the natural statistic for making an inference about  $\alpha_{k-1}$ , for example, in the presence of  $\alpha_1, \dots, \alpha_{k-2}$ , is  $T_{k-1}$ , conditioned on  $T_{k-2}, \dots, T_1$  and  $S_{ij} = n_{ij} + n_{ji}$ . This conditional distribution is given by

$$p_r (T_{k-1} = t_{k-1} \mid T_1 = t_1, \dots, T_{k-2} = t_{k-2}, S_{ij} = s_{ij}, i < j = 1, \dots, k)$$

$$= \frac{p_r (T_1 = t_1, \dots, T_{k-2} = t_{k-2}, T_{k-1} = t_{k-1} \mid S_{ij} = s_{ij}, i < j = 1, \dots, k)}{p_r (T_1 = t_1, \dots, T_{k-2} = t_{k-2} \mid S_{ij} = s_{ij}, i < j = 1, \dots, k)} \tag{7}$$

The distribution in the denominator of (7) is obtained by summing (6) over all possible  $t_{k-1}$  and is given by

$$\begin{aligned}
 p_r(T_1 = t_1, \dots, T_{k-2} = t_{k-2} | S_{ij} = s_{ij}, i < j = 1, \dots, k) \\
 &= \sum_u \frac{C(t_1, \dots, t_{k-2}, u) \exp[\sum_{i=1}^{k-1} \alpha_i t_i]}{\prod_{i=1}^{k-1} \prod_{j=i+1}^k [1 + \exp(\alpha_i - \alpha_j)]^{s_{ij}}}, \\
 &= \frac{\sum_u C(t_1, \dots, t_{k-2}, u) \exp[\sum_{i=1}^{k-1} \alpha_i t_i]}{\prod_{i=1}^{k-1} \prod_{j=i+1}^k [1 + \exp(\alpha_i - \alpha_j)]^{s_{ij}}}.
 \end{aligned} \tag{8}$$

From (6) and (8) the conditional distribution (7) is obtained and is given by

$$\begin{aligned}
 \Pr(T_{k-1} = t_{k-1} | T_1 = t_1, \dots, T_{k-2} = t_{k-2}, S_{ij} = s_{ij}, i < j = 1, \dots, k) \\
 &= \frac{C(t_1, \dots, t_{k-1}) \exp \sum_{i=1}^{k-1} \alpha_i t_i}{\sum_u C(t_1, \dots, t_{k-2}, u) \exp[\sum_{i=1}^{k-2} \alpha_i t_i + \alpha_{k-1} u]}, \\
 &= \frac{C(t_1, \dots, t_{k-1}) \exp(\alpha_{k-1} t_{k-1})}{\sum_u C(t_1, \dots, t_{k-2}, u) \exp(\alpha_{k-1} u)}
 \end{aligned} \tag{9}$$

where  $u$  is an index ranging over the values taken by  $T_{k-1}$  and  $C(t_1, \dots, t_{k-2}, u)$  is the number of distinct set of values of  $n_{ij}$  ( $i < j = 1, \dots, k$ ) which when substituted in (9) yield  $T_1 = t_1, \dots, T_{k-2} = t_{k-2}, T_{k-1} = u$  and  $S_{ij} = s_{ij}$ . Note that (9) does not involve  $\alpha_1, \dots, \alpha_{k-2}$ . In order to simplify the notation, denote  $(t_1, \dots, t_{k-2})$  by  $\underline{t}_{k-2}$ , thus the distribution (9) can be written as

$$p_r(t_{k-1}; \alpha_{k-1}) = \frac{C(\underline{t}_{k-2}, t_{k-1}) \exp(\alpha_{k-1} t_{k-1})}{\sum_u C(\underline{t}_{k-2}, u) \exp(\alpha_{k-1} u)}. \tag{10}$$

An important case of (10) corresponds to  $\alpha_{k-1} = 0$ ,

$$p_r(t_{k-1}; 0) = \frac{C(\underline{t}_{k-2}, t_{k-1})}{\sum_u C(\underline{t}_{k-2}, u)} \tag{11}$$

so that the distribution is determined by the combinatorial coefficients. The computation of the combinatorial coefficient  $C(\underline{t}_{k-2}, t_{k-1})$  involves calculations which are computationally prohibitive for larger value of  $k$ , the number of levels of exposure, and with small numbers of discordant pairs in the resulting cells.

### The Computational Method

A computational method that can be used for obtaining the combinatorial coefficients involved in the distribution (10) is available. Here, interest lies in computing the coefficients  $C(\underline{t}_{k-2}, \cdot)$ , where the dot indicates that the corresponding argument varies over its permissible range of values for  $t_{k-1}$ . The coefficient  $C(\underline{t}_{k-2}, t_{k-1})$  may be counted following the procedure involved in investigating the solutions of the Diophantine systems in non-negative integers as described in Constantine (1987). The Diophantine system is represented by

$$\sum_{r=1}^n a_{ir} x_r = t_i, i = 1, \dots, k-1, \tag{12}$$

where  $a_{ir}$  and  $t_i$  are non-negative integers. Writing

$$\begin{aligned}
 \underline{x} &= (x_1, \dots, x_n), \\
 \underline{t} &= (t_1, \dots, t_{k-1}), \\
 \underline{\xi} &= (\xi_1, \dots, \xi_{k-1})
 \end{aligned}$$

and

$$\underline{\xi}^t = \prod_{i=1}^{k-1} \xi_i^{t_i}.$$

If  $C(\underline{t})$  is the number of solutions to (12), then using the generating function results in

$$\sum_{t \geq 0} C(t) \underline{\xi}^t = \prod_{r=1}^n (1 - \xi_1^{a_{1r}}, \dots, \xi_{k-1}^{a_{(k-1)r}})^{-1}$$

which is

$$\left[ \sum_{t \geq 0} C(t) \underline{\xi}^t \right] \left[ \prod_{r=1}^n (1 - \xi_1^{a_{1r}}, \dots, \xi_{k-1}^{a_{(k-1)r}}) \right] = 1, \tag{13}$$

Equating the coefficients of  $\underline{\xi}^t$  on both sides results in  $C(t)$ , which is the value of the combinatorial coefficient  $C(t_{k-2}, t_{k-1})$ . Note that one of the considerations for computing the coefficients using Diophantine systems is that the  $a_{ir}$ 's and  $t_i$  are non-negative integers valued with non-zero entry in each column of  $(a_{ir})$ . If necessary, this may be achieved by linear transformation with no effect on inference. The non-negativity of the entities involved insures an almost a finite number of solutions to system (12).

The Case of Three Level Exposure

In the simplest situation the polytomous outcome may be considered with three exposure levels. In this case  $k = 3$  with two sufficient statistics from (5) which are:

$$t_1 = n_{12} + 2n_{13} + n_{23} \tag{14}$$

and

$$t_2 = -n_{12} + n_{13} + 2n_{23}. \tag{15}$$

If it is of interest to obtain the distribution of the sufficient statistic  $T_2$  given the observed value of  $T_1 = t_1$ , then the relation (10) reduces to

$$\Pr(t_2; \alpha_2) = \frac{C(t_1, t_2) \exp(\alpha_2 t_2)}{\sum_u C(t_1, u) \exp(\alpha_2 u)}, \tag{16}$$

where  $u$  is an index ranging over all possible values taken by  $T_2$  for given  $T_1 = t_1$  and  $C(t_1, u)$  is the number of distinct set of values for  $n_{12}$ ,  $n_{13}$  and  $n_{23}$  which, when substituted in (14) and (15),

yield  $T_1 = t_1$  and  $T_2 = u$ . The range of  $u$  is determined by considering the maximum and minimum values of  $T_2$ , namely the maximum value of  $t_2 = S_{13} + 2S_{23}$ ; the minimum value of  $t_2$  may be considered to be zero.

The combinatorial coefficients involved in (16) are computed using the method for solutions of the Diophantine system of equations (12). In (15), to insure that  $t_2$  is a positive integer, a linear transformation namely,  $t_2^* = 2t_1 + t_2$ , is considered and provides the Diophantine system of equations:

$$t_1 = x_1 + 2x_2 + x_3 \tag{17}$$

and

$$t_2^* = x_1 + 5x_2 + 4x_3 \tag{18}$$

where  $x_1 = n_{12}$ ,  $x_2 = n_{13}$  and  $x_3 = n_{23}$  respectively. The  $(a_{ir})$  matrix

$$(a_{ir}) = \begin{pmatrix} 1 & 2 & 1 \\ 1 & 5 & 4 \end{pmatrix}$$

results from relations (17) and (18).

Writing  $x = (x_1, x_2, x_3)$ ,  $t^* = (t_1, t_2^*)$ ,  $\underline{\xi} = (\xi_1, \xi_2)$  and inserting the values of  $(a_{ir})$  in (13), the generating function reduces to

$$\left[ \sum_{t^* \geq 0} C(t_1, t_2^*) \xi_1^{t_1} \xi_2^{t_2^*} \right] \begin{bmatrix} 1 - \xi_1 \xi_2 + \xi_1^3 \xi_2^6 \\ -\xi_1 \xi_2^4 + \xi_1^3 \xi_2^9 \\ -\xi_1^4 \xi_2^{10} \end{bmatrix} = 1. \tag{19}$$

The combinatorial coefficient  $C(t_1, t_2^*)$  is obtained by equating the coefficients of  $\xi_1^{t_1} \xi_2^{t_2^*}$  on both sides of (19). This provided the recurrence relation

$$\begin{aligned} & C(t_1, t_2^*) - C(t_1 - 1, t_2^* - 1) \\ & + C(t_1 - 3, t_2^* - 6) - C(t_1 - 1, t_2^* - 4) \\ & + C(t_1 - 3, t_2^* - 9) - C(t_1 - 4, t_2^* - 10) = 0 \end{aligned} \tag{20}$$

The method is repeated for obtaining the values of  $c(t_1, u)$ , where  $0 \leq u \leq \max(t_2^*)$ .

Testing and Estimation

Consider the problem of testing the null hypothesis  $H_0: \alpha_{k-1} = \alpha_{k-1}^0$  against the one-sided alternative  $H_+: \alpha_{k-1} > \alpha_{k-1}^0$ . If  $\underline{t} = (t_1, \dots, t_{k-1})^T$  is the observed vector of sufficient statistics, then following Lehmann (1959) the p-value for the uniformly most powerful unbiased test of  $H_0$  against  $H_+$  is obtained using relation (10) and is given by

$$\begin{aligned}
 p_+(t_{k-1}; \alpha_{k-1}^0) &= p_r(T_{k-1} \geq t_{k-1} | T_1 = t_1, \dots, T_{k-2} = t_{k-2}, S_{ij} = s_{ij}, i < j = 1, \dots, k) \\
 &= p_r(T_{k-1} \geq t_{k-1} | T_{k-2} = t_{k-2}, S_{ij} = s_{ij}, i < j = 1, \dots, k) \\
 &= \sum_{v \geq t_{k-1}} \frac{C(t_{k-2}, v) \exp(\alpha_{k-1}^0 v)}{\sum_u C(t_{k-2}, u) \exp(\alpha_{k-1}^0 u)} \\
 &= \frac{\sum_{v \geq t_{k-1}} C(t_{k-2}, v) \exp(\alpha_{k-1}^0 v)}{\sum_u C(t_{k-2}, u) \exp(\alpha_{k-1}^0 u)}
 \end{aligned}
 \tag{21}$$

Similarly, the p-value for the test of  $H_0$  versus  $H_-: \alpha_{k-1} < \alpha_{k-1}^0$  is given by

$$\begin{aligned}
 p_-(t_{k-1}; \alpha_{k-1}^0) &= \text{pr}(T_{k-1} \leq t_{k-1} | T_{k-2} = t_{k-2}, S_{ij} = s_{ij}, i < j = 1, \dots, k) \\
 &= \frac{\sum_{v \leq t_{k-1}} C(t_{k-2}, v) \exp(\alpha_{k-1}^0 v)}{\sum_u C(t_{k-2}, u) \exp(\alpha_{k-1}^0 u)}
 \end{aligned}
 \tag{22}$$

According to Cox (1970), the upper  $(1-\alpha)$  level confidence limit for  $\alpha_{k-1}$  corresponding to the observed value  $t_{k-1}$  is the solution  $\alpha_{k-1}^U$  to  $P_-(t_{k-1}; \alpha_{k-1}^U) = \alpha$  in (22). Similarly, the solution  $\alpha_{k-1}^L$  to  $P_+(t_{k-1}; \alpha_{k-1}^L) = \alpha$  in (21) gives the lower  $(1-\alpha)$  level confidence limit for  $\alpha_{k-1}$ . These values are evaluated by solving the equations

numerically. The two sided alternatives may be obtained by

$$\begin{aligned}
 p_r(t_{k-1}; \alpha_{k-1}^0) &= \\
 &2 \text{ Min} \left[ P_+(t_{k-1}; \alpha_{k-1}^0), P_-(t_{k-1}; \alpha_{k-1}^0) \right].
 \end{aligned}$$

Tritchler (1984) suggested that the point estimation of the parameter  $\alpha_{k-1}$  denoted by  $\hat{\alpha}_{k-1}$  is the value which nearly satisfies  $P_-(t_{k-1}; \hat{\alpha}_{k-1}) = P_+(t_{k-1}; \hat{\alpha}_{k-1}) = 0.5$ . Following similar techniques, the point and interval estimation of the other model parameters can be obtained.

Numerical Example

The methodology of exact analysis described herein is most suitably performed on a computer, however, for illustration purposes, consider a hypothetical matched pairs data set involving a response variable taking three levels as shown in Table 2. The estimates and 95 percent confidence limits of the parameters for the data set shown in Table 2 were obtained using the exact method of analysis; results are presented in Table 3.

Table 2: Frequency Distribution of Case-Control Pairs with Three Exposure Levels

Exposure Level for Case (F <sub>1</sub> )	Exposure Level for Control (F <sub>0</sub> )		
	1	2	3
1	20	3	2
2	2	15	2
3	3	2	10

Table 3: Results of Logistic Analysis of the Data in Table 2 Based on Exact Method

Parameters	Exact Estimate	95 Percent Confidence Limits
$\alpha_1$	-0.120	(-0.630, 0.270)
$\alpha_2$	-0.150	(-1.050, 0.350)
$\alpha_3$	0.270	(-0.620, 1.680)

## EXACT LOGISTIC REGRESSION FOR A MATCHED PAIRS CASE-CONTROL DESIGN

The data set used in the example was considerably large with 59 matched case-control pairs; however, the number of discordant pairs of observations involved in the analysis was very small. Hence, in this situation, the exact method of analysis may be more appropriate. This article considered an exact method of analysis in case of 1-1 matched case-control data when the risk factor of interest is polytomous in nature.

### References

- Agresti, A. (1990). *Categorical data analysis*. New York, NY: J. Wiley and Sons.
- Breslow, N. E., & Day N. E. (1980). *Statistical methods in cancer research, Vol. 1: The analysis of case and control studies*. Lyon: International Agency for Research on Cancer.
- Cox, D. R. (1970). *Analysis of binary data*. Mathuen: London.
- Constantine, G. M. (1987). *Combinatorial theory and statistical design*. New York, NY: J. Wiley and Sons.
- Dubin, N., & Pasternack, B. S. (1986). Risk assessment for case-control subgroups by polytomous logistic regression. *American Journal of Epidemiology*, 123, 1101-1117.
- Ejigou, A., & McHugh, R. (1977). Estimation of relative risk from matched pairs in epidemiologic research. *Biometrics*, 33, 552-558.
- Ejigou, A., & McHugh, R. (1981). Relative risk estimation under multiple matching. *Biometrika*, 68, 85-91.
- Ganguly, S. S., & Naik-Nimbalkar, U. (1992). Use of polytomous logistic model in matched case-control studies. *Biometrical Journal*, 34, 209-217.
- Ganguly, S. S., & Naik-Nimbalkar, U. (1992). Polytomous logistic model for matched case-control design. *Journal of Applied Statistics*, 19(4), 455-464.
- Ganguly, S. S., & Naik-Nimbalkar, U. V. (1995). Analysis of ordinal data in a study of endometrial cancer under a matched pairs case-control design. *Statistics in Medicine*, 14, 1545-1552.
- Ganguly S. S. (2006). Cumulative logit models for matched pairs case-control design: Studies with covariates. *Journal of Applied Statistics*, 33(5), 513-522.
- Hirji, K. F. (1992). Computing exact distributions for polytomous response data. *Journal of the American Statistical Association*, 87, 487-492.
- Hirji, K. F, Mehta, C. R., & Patel, N. R. (1988). Exact inference for matched case-control studies. *Biometrics*, 44, 803-814.
- Holford, T. R. (1978). The analysis of pair-matched case-control studies, a multivariate approach, *Biometrics*, 34, 665-672.
- Holford, T. R., While, C., & Kelsey, J. L. (1978). Multivariate analysis for matched case-control studies. *American Journal of Epidemiology*, 107, 246-256.
- Kleinbaum, D. G., Kupper, L. L., & Chambless, L. E. (1982). Logistic regression analysis of epidemiologic data: Theory and practice. *Communications in Statistics - Theory and Methods*, 11(5), 485-547.
- Lehmann, E. L. (1959). *Testing statistical hypothesis*. New York, NY: J. Wiley and Sons.
- Liang, K. Y., & Stewart, W. F. (1987). Polytomous logistic regression for matched case-control studies with multiple case or control groups. *American Journal of Epidemiology*, 125, 720-730.
- Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of National Cancer Institute*, 22, 719-749.
- Mehta, C. R. (1994). The exact analysis of contingency tables in medical research. *Statistical Methods in Medical Research*, 3, 135-156.
- Miettinen, O. (1970). Estimation of relative risk from individually matched series. *Biometrics*, 26, 75-86.
- Pagano, M., & Tritchler, D. (1983). On obtaining permutation distributions in polynomial time. *Journal of the American Statistical Association*, 78, 435-440.
- Pike, M. C., Casagrande, J., & Smith, P. G. (1975). Statistical analysis of individually matched case-control studies in epidemiology: factor under study a discrete variable taking multiple values. *British Journal of Preventive and Social Medicine*, 29, 196-201.
- Prentice, R. (1976). Use of the logistic model in retrospective studies. *Biometrics*, 32, 599-606.
- Prentice, R. L., & Pyke, R. (1979). Logistic disease incidence models and case-control studies. *Biometrics*, 66, 403-411.
- Tritctler, D (1984). An algorithm for exact logistic regression. *Journal of the American Statistical Association*, 79, 709-711.