Journal of Biometrics & Biostatistics
ISSN: 2155-6180

Research Article Open Access

Statistical Analysis of Case-Control Data of Endometrial Cancer Based on New Asymmetry Models

Kouji Yamamoto[1]* and Sadao Tomizawa[2]

[1]*Center for Clinical Investigation and Research, Osaka University Hospital, 2-15, Yamadaoka, Suita, Osaka, 565-0871, Japan*
[2]*Department of Information Sciences, Faculty of Science and Technology,Tokyo University of Science, Noda City, Chiba, 278-8510, Japan*

Abstract

Background: For the data from the Los Angeles study in Breslow and Day of endometrial cancer and obtained from the 59 matched pairs using four dose levels of conjugated oestrogen, this study proposes new statistical models and gives an easy interpretation, as an approach to assess the data more properly.

Methods: Proposing new statistical models for analyzing the endometrial cancer data, we apply them to the data, compare and assess the models considered here.

Results: We have found a more preferable model which fits the data better than some existing models. Under the preferable model, we have seen that the average dose of oestrogen for case in a matched pair tends to be more than that for control in the pair.

Conclusions: We have proposed two kinds of statistical models and made a conclusion that average dose for case tends to be more than that for control.

Keywords: Asymmetry; Cumulative linear diagonals-parameter; Endometrial cancer; Model; Square contingency table

Introduction

Consider the data in Table 1 taken directly from Breslow and Day [1] . These are from the Los Angeles study of endometrial cancer and obtained from the 59 matched pairs using four dose levels of conjugated oestrogen, (1) none, (2) 0.1-0.299 mg, (3) 0.3-0.625 mg, and (4) 0.626+mg (/day). For these data, we are interested in (a) what times the probability that the average dose of oestrogen for case in a matched pair is in category i and that for control in the pair is in category j ($<i$) is higher than the probability that the average dose for case in the pair is in category j and that for control is in category i ($>j$), and (b) what times the probability that the average dose for case in a pair is in category i or above and that for control in the pair is in category j ($<i$) or below is higher than the probability that the average dose for case in the pair is in category j or below and that for control is in category i ($>j$) or above. Especially, we are interested in what times the probability that the average dose for case in a pair is not zero (i.e., in categories 2, 3, and 4) and that for control in the pair is zero (i.e., in category 1) is higher than the probability that the average dose for case in the pair is zero and that for control is not zero. Namely we are interested in seeing what structure of asymmetry for probabilities there is between the average dose for case in a pair and that for control in the pair.

Agresti considered an asymmetry model, called the linear diagonals-parameter symmetry (LDPS) model [2]. Miyamoto et al. considered an asymmetry model, called the cumulative linear diagonals-parameter symmetry (CLDPS) model [3] , and applied this model to the data in Table 1.

The present paper (1) reviews some asymmetry models, (2) proposes new asymmetry models which are generalizations of the LDPS model and CLDPS model, and (3) analyzes the data in Table 1 using these new models.

Material and Methods

Reviews of models

Consider an $r \times r$ square contingency table with the same row and column classifications, as Table 1. Let p_{ij} denote the probability that an observation will fall in the i th row and jth column of the table ($i = 1,\dots,r; j = 1,\dots,r$). As a model which indicates the structure of asymmetry for $\{p_{ij}\}$, the LDPS model is given as

$$\frac{p_{ij}}{p_{ji}} = \delta^{i-j} \quad (i > j).$$

For the endometrial cancer data in Table 1, this model indicates that the probability that the average dose of oestrogen for case in a matched pair is in category i and that for control in the pair is in category j ($<i$) is δ^{i-j} times higher than the probability that the average dose for case in the pair is in category j and that for control is in category i ($> j$). If $\delta > 1$, then the average dose of oestrogen for case in a pair tends to be more than that for control in the pair. A special case of the LDPS model obtained by putting $\delta = 1$ is the symmetry (S) model [4,5]. Also the LDPS model with $\{\delta^{i-j}\}$ replaced by $\{\gamma\delta^{i-j}\}$ is the two ratio-parameter symmetry (2RPS) model [6]. A special case of the 2RPS model obtained by putting $\delta = 1$ is the conditional symmetry (CS) model [7].

Let for $i > j$,

$$G_{ij} = \sum_{s=i}^{r}\sum_{t=1}^{j} p_{st} \quad \text{and} \quad G_{ji} = \sum_{s=1}^{j}\sum_{t=i}^{r} p_{st}.$$

For the endometrial cancer data, (1) G_{ij} for $i > j$ indicates that the cumulative probability that the average dose for case in a pair is in category i or above and that for control in the pair is in category j or

***Corresponding author:** Kouji Yamamoto, Center for Clinical Investigation and Research, Osaka University Hospital, 2-15, Yamadaoka, Suita, Osaka, 565-0871, Japan, E-mail: yamamoto-k@hp-crc.med.osaka-u.ac.jp

Received May 07, 2012; **Accepted** June 19, 2012; **Published** June 21, 2012

Citation: Yamamoto K, Tomizawa S (2012) Statistical Analysis of Case-Control Data of Endometrial Cancer Based on New Asymmetry Models. J Biomet Biostat 3:147. doi:10.4172/2155-6180.1000147

Average dose	Average dose for control (mg/day)				
for case	0	0.1-0.299	0.3-0.625	0.626+	Total
(mg/day)	(1)	(2)	(3)	(4)	
0 (1)	6	2	3	1	12
	(6.00)	(3.33)	(2.04)	(1.23)	
0.1-0.299 (2)	9	4	2	1	16
	(8.27)	(4.00)	(1.48)	(0.73)	
0.3-0.625 (3)	9	2	3	1	15
	(9.66)	(2.17)	(3.00)	(1.32)	
0.626+ (4)	12	1	2	1	16
	(11.80)	(1.07)	(1.90)	(1.00)	
Total	36	9	10	4	59

Table 1: Average doses of conjugated oestrogen used by cases and matched control: Los Angeles endometrial cancer study [1] (The parenthesized values are maximum likelihood estimates of expected frequencies under the CLDPS(3) model).

below, and (2) G_{ji} for $i > j$ indicates that the cumulative probability that the average dose for case in a pair is in category *j* or below and that for control in the pair is in category *i* or above.

As a model which indicates the structure of asymmetry for $\{G_{ij}\}$, $i \neq j$, the CLDPS model is defined by

$$\frac{G_{ij}}{G_{ji}} = \Delta^{i-j} \quad (i > j).$$

The CLDPS model is different from the LDPS model. For the endometrial cancer data in Table 1, the CLDPS model indicates that the probability that the average dose for case in a pair is in category *i* or above and that for control in the pair is in category j $(< i)$ or below is Δ^{i-j} times higher than the probability that the average dose for case in the pair is in category *j* or below and that for control is in category *i* or above. If $\Delta > 1$, then the average dose for case in a pair tends to be more than that for control in the pair. Also the CLDPS model with $\{\Delta^{i-j}\}$ replaced by $\{\Gamma\Delta^{i-j}\}$ is the cumulative two ratios-parameter symmetry (C2RPS) model [8].

New models

We shall propose two kinds of new models. First, consider a generalization of the LDPS model as follows: for a fixed K(K=0,1,2,..;-1,-2,…),

$$\frac{p_{ij}}{p_{ji}} = \delta^{K+(i-j)} \quad (i > j).$$

We shall denote this model by LDPS(K). Then the LDPS(0) model is equivalent to the LDPS model. Also the LDPS $(-r)$ model is equivalent to another LDPS model, proposed by Tomizawa [9]. For the endometrial cancer data, the LDPS(K) model indicates that the probability that the average dose of oestrogen for case in a pair is in category *i* and that for control in the pair is in category $j(<i)$ is $\delta^{K+(i-j)}$ times higher than the probability that the average dose for case in the pair is in category *j* and that for control is in category $i(>j)$. If $\delta > 1$ with $K \geq 1$, then the average dose for case in a pair tends to be more than that for control in the pair, and the tendency is stronger under the LDPS(K) model than under the LDPS model, because $\delta^{K+(i-j)} > \delta^{i-j} > 1$ with $\delta > 1$, $K \geq 1$, and $i>j$.

Secondly, consider a generalization of the CLDPS model as follows: for a fixed K (K = 0,1,2,..;-1,-2,…),

$$\frac{G_{ij}}{G_{ji}} = \Delta^{K+(i-j)} \quad (i > j).$$

We shall denote this model by CLDPS(K). Then the CLDPS(0) model is equivalent to the CLDPS model. For the endometrial cancer data, the CLDPS(K) model indicates that the probability that the average dose of oestrogen for case in a pair is in category *i* or above and that for control in the pair is in category $j(<i)$ or below is $\Delta^{K+(i-j)}$ times higher than the probability that the average dose for case in the pair is in category *j* or below and that for control is in category *i* or above. If $\Delta>1$ with $K \geq 1$, then the average dose for case in a pair tends to be more than that for control in the pair, and the tendency is stronger under the CLDPS(K) model than under the CLDPS model, because $\Delta^{K+(i-j)} > \Delta^{i-j} > 1$ with $\Delta>1$, $K \geq 1$, and $i>j$.

The CLDPS(K) model is different from the LDPS(K) model. The CLDPS(K) model indicates how the cumulative probabilities $\{G_{ij}\}$ for $i>j$ are asymmetric to $\{G_{ij}\}$, and the LDPS model indicates how the cell probabilities $\{P_{ij}\}$ for $i>j$ are asymmetric to $\{P_{ij}\}$. For the endometrial cancer data, we are also interested in seeing what times the probability that the average dose of oestrogen for case in a pair is not zero (i.e., in categories 2, 3, and 4) and that for control in the pair is zero (i.e., in category 1) is higher than the probability that the average dose for case in the pair is zero and that for control is not zero. We can see under the CLDPS(K) model that the probability that the average dose for case in a pair is not zero and that for control in the pair is zero is $\Delta^{K+1}(= G_{21}/G_{12})$ times higher than the probability that the average dose for case in the pair is zero and that for control is not zero, although we cannot see such a structure under the LDPS(K) model.

Test of goodness-of-fit of model

Let n_{ij} denote the observed frequency in the (i,j)th cell of the $r \times r$ table $(i=1,\ldots,r; j=1,\ldots,r)$, with $n=\sum\sum n_{ij}$, and let m_{ij} denote the corresponding expected frequency. Assume that $\{n_{ij}\}$ have a multinomial distribution. The maximum likelihood estimates of expected frequencies $\{m_{ij}\}$ under each model could be obtained, for example, using the Newton-Raphson method to the log-likelihood equations. Each model can be tested for goodness-of-fit by e.g., the likelihood ratio chi-squared statistic G^2 with the corresponding degrees of freedom, defined by

$$G^2 = 2\sum_{i=1}^{r}\sum_{j=1}^{r} n_{ij} \log\left(\frac{n_{ij}}{\hat{m}_{ij}}\right),$$

where $\hat{m}_{ij}$ is the maximum likelihood estimate of m_{ij} under the model. The numbers of degrees of freedom for the LDPS(K) and CLDPS(K) models are both $(r+1)(r-2)/2$, which is one less than that for the S model and one more than that for the 2RPS (C2RPS) model.

Analysis of Data

We shall analyze the endometrial cancer data in Table 1 using the models in above section. Table 2 gives the values of likelihood ratio test statistic G^2 for each model. Note that the LDPS(0) model is equivalent to the LDPS model, and the CLDPS(0) model is equivalent to the CLDPS model.

The S model fits these data poorly. Therefore it is estimated that the probability that the average dose of oestrogen for case in a matched pair is in category i and that for control in the pair is in category $j(<i)$ is not equal to the probability that the average dose for case in the pair is in category j and that for control is in category $i(>j)$.

Among the LDPS(K) models for various K, the LDPS(0) model (i.e., the LDPS model) provides the best-fitting with 5 degrees of freedom, which fits better than the CS model with same 5 degrees of freedom (Table 2).

Also, the LDPS(0) model is a special case of the 2RPS model, obtained by putting $\gamma = 1$. Since the 2RPS model fits these data well, we shall test the hypothesis of $\gamma = 1$ (i.e., the hypothesis that the LDPS(0) model holds) under the assumption that the 2RPS model holds. It can be tested according to the difference between the likelihood ratio statistic G^2 for the LDPS(0) model and that for the 2RPS model. The difference is 0.06 with 1 degree of freedom. Therefore we can accept the hypothesis of $\gamma = 1$ in the 2RPS model, at the 0.05 level ($p = 0.806$). Thus the LDPS(0) model would be preferable to the 2RPS model for these data.

Next, among the CLDPS(K) models for various K, the CLDPS(3) model provides the best-fitting with 5 degrees of freedom (Table 2). The CLDPS(3) model fits these data better than the CLDPS(0) model with both 5 degrees of freedom.

Also, the CLDPS(3) model is a special case of the C2RPS model, obtained by putting $\Gamma = \Delta^3$. Since the C2RPS model fits these data well, we shall test the hypothesis of $\Gamma = \Delta^3$ (i.e., the hypothesis that the CLDPS(3) model holds) under the assumption that the C2RPS model holds. The difference between the likelihood ratio statistic G^2 for the CLDPS(3) model and G^2 for the C2RPS model is 0.02 with 1 degree of freedom. Therefore we can accept the hypothesis of $\Gamma = \Delta^3$ in the C2RPS model, at the 0.05 level (p = 0.888). Thus the CLDPS(3) model is preferable to the C2RPS model for these data. Therefore for the endometrial cancer data in Table 1, the CLDPS(3) model is the best-fitting model among the models given in Table 2.

Models	Degrees of freedom	G^2	p -value
S	6	19.27**	0.004
CS	5	4.56	0.472
2RPS	4	2.91	0.572
C2RPS	4	1.52	0.823
LDPS(-4)	5	9.16	0.103
LDPS(-3)	5	13.40*	0.020
LDPS(-2)	5	18.98**	0.002
LDPS(-1)	5	5.81	0.325
LDPS(0)	5	2.97	0.704
LDPS(1)	5	3.00	0.701
LDPS(2)	5	3.21	0.668
LDPS(3)	5	3.40	0.639
LDPS(4)	5	3.55	0.616
CLDPS(-4)	5	13.14*	0.022
CLDPS(-3)	5	15.23**	0.009
CLDPS(-2)	5	17.74**	0.003
CLDPS(-1)	5	19.27**	0.002
CLDPS(0)	5	9.85	0.080
CLDPS(1)	5	3.42	0.636
CLDPS(2)	5	1.89	0.865
CLDPS(3)	5	1.54	0.909
CLDPS(4)	5	1.55	0.907

Table 2: Values of likelihood ratio chi-squared statistic G^2 for models applied to the data in Table 1. (The symbols * and ** mean significant at the 0.05 and 0.01 levels, respectively).

Under the CLDPS(3) model applied to these data, the maximum likelihood estimate of Δ is $\hat{\Delta} = 1.457$. Thus the maximum likelihood estimates of $\{\Delta^{3+(i-j)}\}$, $i-j = 1,2,3,$ are $\hat{\Delta}^4 = 4.502$, $\hat{\Delta}^5 = 6.558$, and $\hat{\Delta}^6 = 9.552$. Hence, under the CLDPS(3) model, the probability that the average dose of oestrogen for case in a matched pair is in category i or above and that for control in the pair is in category $j(<i)$ or below is estimated to be $\hat{\Delta}^{3+(i-j)}$ times higher than the probability that the average dose for case in the pair is in category j or below and that for control is in category i or above.

Especially, under the CLDPS(3) model, the probability that the average dose for case in a pair is not zero (i.e., in categories 2, 3, and 4) and that for control in the pair is zero (i.e., in category 1) is estimated to be 4.502 ($= \hat{\Delta}^4$) times higher than the probability that the average dose for case in the pair is zero and that for control is not zero. Also under the CLDPS(3) model, the probability that the average dose for case in a pair is 0.626+ (mg/day) (i.e., in category 4) and that for control in the pair is zero (i.e., in category 1) is estimated to be 9.552 ($= \hat{\Delta}^6$) times higher than the probability that the average dose for case in the pair is zero and that for control is 0.626+ (mg/day).

Since $\hat{\Delta}^{3+(i-j)} > 1$ for $i>j$, under the CLDPS(3) model it is estimated that the average dose for case in a pair tends to be more than that for control in the pair.

Discussion

For the endometrial cancer data in Table 1, we shall discuss why the CLDPS(3) model fits better than the CLDPS(0) model (i.e., the CLDPS model). Note that the CLDPS(K) model is a special case of the C2RPS model, obtained by putting $\Gamma = \Delta^K$. For the endometrial cancer data, the maximum likelihood estimates of parameters Γ and Δ under the C2RPS model are $\hat{\Gamma} = 3.194$ and $\hat{\Delta} = 1.410$. Thus it seems that $\hat{\Gamma}$ is close to $\hat{\Delta}^3 = 2.803$. This would show that the CLDPS(3) model fits the endometrial cancer data well.

Conclusions

We have proposed two kinds of asymmetry models, namely, the LDPS(K) model and the CLDPS(K) model. The LDPS(K) model is useful for seeing the structure of asymmetry of cell probabilities $\{P_{ij}\}$, and the CLDPS(K) model is useful for seeing the structure of asymmetry of cumulative probabilities $\{G_{ij}\}$.

For the endometrial cancer data in Table 1, we have seen using the CLDPS(3) model that the average dose of oestrogen for case in a matched pair tends to be more than that for control in the pair; especially, there is the structure of strong asymmetry such that the probability that the average dose for case in a pair is 0.626+ (mg/day) and that for control in the pair is zero is 9.552 times higher than the probability that the average dose for case in the pair is zero and that for control is 0.626+ (mg/day).

Acknowledgements

The authors would like to thank the editor and an anonymous reviewer for their helpful comments and suggestions.

References

1. Breslow NE, Day NE (1980) Statistical Methods in Cancer Research, Vol. I-The Analysis of Case-Control Studies. International Agency for Research on Cancer Scientific Publications, Lyon, France 1: 338.
2. Agresti A (1983) A simple diagonals-parameter symmetry and quasi-symmetry model. Statistics and Probability Letters 1: 313-316.
3. Miyamoto N, Ohtsuka W, Tomizawa S (2004) Linear diagonals-parameter symmetry and quasi-symmetry models for cumulative probabilities in square contingency tables with ordered categories. Biometrical Journal 46: 664-674.
4. Bowker AH (1948) A test for symmetry in contingency tables. Journal of the American Statistical Association 43: 572-574.
5. Bishop YMM, Fienberg SE, Holland PW (1975) Discrete Multivariate Analysis: Theory and Practice. The MIT Press, Cambridge, Massachusetts.
6. Tomizawa S (1987) Decompositions for 2-ratios-parameter symmetry model in square contingency tables with ordered categories. Biometrical Journal 29: 45-55.
7. McCullagh P (1978) A class of parametric models for the analysis of square contingency tables with ordered categories. Biometrika 65: 413-418.
8. Tomizawa S, Miyamoto N, Yamamoto K, Sugiyama A (2007) Extensions of linear diagonal-parameter symmetry and quasi-symmetry models for cumulative probabilities in square contingency tables. Statistica Neerlandica 61: 273-283.
9. Tomizawa S (1990) Another linear diagonals-parameter symmetry model for square contingency tables with ordered categories. South African Statistical Journal 24: 117-125.