Acta Genet Med Gemellol 30:267-274 (1981)
The Mendel Institute/Alan R. Liss, Inc.

Received 16 July 1981

Final 27 November 1981

Multivariate Logit Analysis of Concordance Ratios for Qualitative Traits in Twin Studies

Jaakko Kaprio, Seppo Sarna, Markku Koskenvuo

Department of Public Health Science, University of Helsinki

The application of multiway contingency table analysis to the multivariate analysis of concordance ratios in twin studies is developed. The approach is illustrated by data on smoking and alcohol use in Finland and Sweden. This approach can enable the assessment of the effect of other variables on the concordance ratio and thus allow estimates of genetic effects on the trait under study. Hypotheses on relationships between genetic effects and other variables can be tested. After hypothesis testing, model fitting of the best hypothesis can be carried out.

Key words: Twin concordance, Contingency tables, Multivariate analysis, Smoking, Alcohol

INTRODUCTION

A usual way of analysing qualitative twin data (eg, disease versus nondisease state) is to compute concordance ratios. These may be expressed as pairwise ratios, probandwise or as coincidence ratios depending on the study design, method of case ascertainment, and trait under study [2,8,18]. Commonly the concordance ratio is expressed for a single trait in one sample. Comparisons of concordance ratio are carried out between monozygotic (MZ) and dizygotic (DZ) pairs; higher concordance ratios in MZ than in DZ pairs are interpreted as evidence in favor of a genetic effect on the trait, though in many cases MZ twins may share more environmental influences as well.

The expression of traits, however, is often dependent on many factors in the environment, and therefore analysis of the effect of environmental variables on genetic factors is of interest when studying various processes and interactions. In epidemiological studies, for example, we are interested in how risk factors are related to disease, and in how their effect is modulated by modifying factors. Genetic factors can influence various parts of these processes, and specification of which parts and how is in the interest of genetic epidemiology.

This study has been supported by a grant from the Council for Tobacco Research, USA, Inc.

https://doi.org/10.1017/S0001566000006474 Published online by Cambridge University Press

In the study of the effect of various variables, relationships between different variables are termed interactions. In twin studies one variable is, of course, zygosity. Depending on the study design other variables may be used, eg, sex, age group, etc. For such qualitative traits the proportion of twin pairs that are concordant for the study trait within the category of the variable is often studied. Thus, it may be of interest to study whether the proportion of trait concordant pairs is similar in MZ and DZ pairs of both sexes. Because of multiway interactions there is often a need to analyze three or more variables simultaneously. A frequent form of presentation of this type of data is the contingency table, permitting comparisons between groups.

A sophisticated way of treating contingency tables is to use models based on the probability distributions of the exponential type. One of the most frequently used models is the log-linear model [3,7,10,11,14]. Since in twin studies the concordance ratio is of interest, a different but analogous model is better suited, the logit model. This paper will present a new approach for the analysis of certain type of twin data.

DEFINITION OF THE MODEL

Let us suppose that a trait to be studied is measured by the concordance ratio (X). For the purpose of the model we define this variable as a ratio of the number of concordant pairs to the number of nonconcordant pairs. Let us assume that we want to study the association of r qualitative variables with the concordance ratio. There are also s variables that are considered to be confounding or modifying factors in this association, so that their effect has to be controlled by the model. Suppose that the data is arrayed in the form of a p-dimensional (p = r + s) contingency table, whose elements are characterized with the variables:

$$X_{I1,\ldots,Ip} = K_{I1,\ldots,Ip} / (N_{I1,\ldots,Ip} - K_{I1,\ldots,Ip}), \qquad (1)$$

where indexes I1, . . . , Ip indicate the different categories of the p factors, $K_{I1}, \ldots, K_{Ip}$ are the numbers of concordant pairs, and $N_{I1}, \ldots, N_{Ip}$ the total number of pairs in the different categories.

The approach is illustrated by data from a study on smoking and alcohol use in adult like-sexed twin pairs in Finland and Sweden [15]. Smoking habits of 19,793 adult like-sexed twin pairs in Finland and Sweden were clarified by postal questionnaire surveys. Current cigarette smokers were defined as the persons who had never smoked more than 5–10 packs of cigarettes and who presently smoked regularly. A smoking concordant pair was defined as a pair in which both twins were current cigarette smokers. The distribution of smoking concordant pairs was categorized by zygosity (Z), sex (S), and country (C) (Table 1). Here p = 3, and since zygosity is the variable of study interest, and sex and country of residence are considered confounders, r = 1 and s = 2. All three p variables are in this case dichotomous. The logit [ln $(X_{I1}, \ldots, {}_{Ip})$] for each cell of this three way classification has been computed in Table 1.

For the analysis of the specified contingency table we define the following saturated model:

$$\ln E(X_{I1,\ldots,Ip}) = \Theta^{(0)} + \sum_{i} \Theta^{(i)}_{Ii} + \sum_{\substack{i \\ i<j}} \sum_{j} \Theta^{(ij)}_{IiIj} + \ldots + \Theta^{(12,\ldots,p)}_{I1,I2,\ldots,Ip} \qquad (2)$$

https://doi.org/10.1017/S0001566000006474 Published online by Cambridge University Press

where

$E(X_{I1,...,Ip})$ is the expected value of the concordance ratio according to the model used,

$\Theta^{(0)}$ is the parameter indicating overall effect,

$\Theta^{(i)}_{Ii}$, i = 1, . . . , p are parameters indicating the main effect of the factors 1, . . . , p,

$\Theta^{(ij)}_{Ii,Ij}$, i,j = 1,2, . . . , p; i < j are second-order interaction terms, and

$$\vdots$$

$\Theta^{(1...p)}_{I1,...,Ip}$ is the pth order interaction term.

This model is called the logit model, and it assumes an underlying binomial probability distribution.

The estimation of the parameters in the logit model is usually carried out by maximum likelihood method. A number of computer programs are available which give maximum-likelihood estimates of all parameters. Algorithms have been published [12,13] and the widely used program package GLIM [5] contains programs that can be used for the analysis of contingency tables.

Hypothesis Testing

We may formulate a number of hypotheses in terms of different Θs in the model [2] being 0. A large number of hypotheses based on the parameters of the model can be formulated. For a simple analysis of three dichotomous traits, there are up to 64 different hypotheses, whereas for a five-way table there are several thousand different hypotheses. One has to limit oneself to groups of hypotheses that have a meaningful interpretation; some hypotheses have interpretations in terms of independence, conditional independence, and uniform distribution [10].

TABLE 1. Number of Smoking Positive Concordant Pairs by Zygosity, Sex, and Country

Country (C) (I3)	Sex (S) (I2)	Zygosity (Z) (I1)	Positively concordant pairs	Total pairs	Logit
Sweden	Male	MZ	452	2011	−1.24
		DZ	600	3204	−1.47
	Female	MZ	641	2413	−1.02
		DZ	839	3705	−1.23
Finland	Male	MZ	373	1210	−0.81
		DZ	727	2899	−1.09
	Female	MZ	232	1430	−1.64
		DZ	469	2921	−1.65

https://doi.org/10.1017/S0001566000006474 Published online by Cambridge University Press

A natural way to start is to compute test statistics for certain basic hypotheses:

$$H_{12\,.\,.\,.\,p} \;:\; \Theta^{(12\,.\,.\,.\,p)}_{I1,I2,\,.\,.\,.\,Ip} = 0 \quad \text{for all } I1,I2,\,.\,.\,.\,,\,Ip$$

$$H_{13\,.\,.\,.\,p} \;:\; \Theta^{(12\,.\,.\,.\,p)}_{I1,I2,\,.\,.\,.\,Ip} = 0 \text{ and } \Theta^{(13\,.\,.\,.\,p)}_{I1,I3,\,.\,.\,.\,,\,Ip} = 0 \quad \text{for all } I1,I2,\,.\,.\,,\,Ip$$

.

.

.

$$H_{12\,.\,.\,.\,p\text{-}1} \;:\; \Theta^{(12\,.\,.\,.\,p)}_{I1,I2,\,.\,.\,.\,Ip} = 0 \text{ and } \Theta^{(12\,.\,.\,.\,p\text{-}1)}_{I1,I3,\,.\,.\,.\,,\,Ip\text{-}1} = 0 \quad \text{for all } I1,I2,\,.\,.\,.\,,\,Ip$$

.

.

.

$$H_1 : \text{all interactions of higher order are 0 and } \Theta^{(1)}_{I1} = 0 \quad \text{for all } I1$$

.

.

.

$$H_p : \text{all int eractions of higher order are 0 and } \Theta^{(p)}_{Ip} = 0 \quad \text{for all } Ip$$

$$H_0 : \text{all parameters are 0 except } \Theta^{(0)}. \tag{3}$$

These hypotheses are tested by the log likelihood statistic, which is asymptotically chi-square distributed [17]. The statistic is denoted by Z(H), and called deviance in GLIM [5].

The sufficient marginals for the hypotheses related to the example data are shown in Table 2. In the same table are the test statistics, the associated degrees of freedom, and the P values for each hypothesis.

In addition to the basic hypotheses shown in Table 2, composite hypotheses can be constructed. These are used in sequential testing, when the effect of variables or interactions, as sole effects, on the concordance ratio is assessed. It is seen that four combinations of sufficient marginals yielded test statistics that did not cause rejection of the corresponding hypotheses (Table 2).

Sequential testing of hypotheses is a basic approach in the analysis of contingency tables [4]. All the hypotheses to be tested are systematically arranged in a hierarchial order, which gives us which hypotheses to test first and which hypotheses are tested under the assumption that other, earlier hypotheses hold.

There are, however, few rules for how to select the sequence of successive hypotheses to be tested. Several alternative criteria for hypothesis selection have been suggested [6]. The level of significance for rejection of a hypothesis will of course be determined by the person setting the hypotheses.

For composite hypotheses a simple formula exists,

TABLE 2. Summary of Sufficient Marginals and Test Statistics for Example Data

Hypothesis	Deviance Z(H)	df	P value	Sufficient marginals zygosity(Z), sex(S), country(C)
H_{ZSC}	2.982	1	0.084	S·C + Z·C + Z·S
H_{SC}	5.660	2	0.059	Z·S + Z·C
H_{ZS}	162.2	2	<0.001	S·C + Z·C
H_{ZC}	3.819	2	0.148	Z·S + S·C
$H_{SC,ZC}$	163.4	3	<0.001	Z·S + C
$H_{ZC,ZS}$	6.156	3	0.104	S·C + Z
$H_{SC,ZS}$	162.5	3	<0.001	Z·C + S
$H_{ZC,ZS,SC}$	163.6	4	<0.001	Z + S + C
H_{C}	165.7	5	<0.001	Z + S
H_{S}	179.4	5	<0.001	Z + C
H_{Z}	192.4	5	<0.001	S + C
$H_{Z,C}$	195.9	6	<0.001	S
$H_{S,C}$	181.2	6	<0.001	Z
$H_{Z,S}$	207.3	6	<0.001	C
$H_{Z,C,S}$	210.5	7	<0.001	—

TABLE 3. Table of Variation of Logit Analysis of Smoking Concordance by Zygosity, Sex, and Country

Variation due to		Hypothesis	Deviance (df = 1)	P value
3rd-order interaction		H_{ZSC}	2.98	0.08
2nd-order interaction	(Z × C)	H_{ZC} ; H_{ZSC}	0.84	0.36
2nd-order interaction	(Z × S)	$H_{ZC,ZS}$; H_{ZC}	2.34	0.13
2nd-order interaction	(S × C)	$H_{ZC,ZS,SC}$; $H_{ZC,ZS}$	157.44	<0.001
Main effect: zygosity		H_{Z} ; $H_{ZC,ZS,SC}$	28.80	<0.001
Main effect: country		$H_{Z,C}$; H_{Z}	3.50	0.06
Main effect: sex		$H_{Z,C,S}$; $H_{Z,C}$	14.60	<0.001

$$Z(H;H^*) = Z(H) - Z(H^*), \tag{4}$$

that enables us to compute the value of the test statistic to test for H against H*, where H* is a more general basic hypothesis that subsumes H. For example, the composite hypothesis for a second-order interaction between zygosity and sex is given by H_{ZC} ; H_{ZSC}. The Z statistic is computed by the formula:

$$Z(H_{ZC} ; H_{ZSC}) = Z(H_{ZC}) - Z(H_{ZSC}),$$

$$0.84 = 3.82 - 2.98.$$

In this way a table of tests for the interaction terms and main variable effects analogous to tables of variation in analysis of variance can be constructed. For the empirical data, the entire table of variation is shown in Table 3.

For the data on cigarette smoking in male and female MZ and DZ twin pairs in Finland and Sweden, the hypotheses were sequentially tested. The hypothesis of no three-way interaction of zygosity, sex, and country is not rejected (P = 0.08) at the P = 0.05

https://doi.org/10.1017/S0001566000006474 Published online by Cambridge University Press

TABLE 4. Model Estimates, Standard Errors, and P Values for Logit Model of Smoking Concordance Ratios

Estimate		Standard			
N:o	Value	error	Parameter	Est/SE	P value
1	−1.256	0.0404	$\Theta^{(0)}$: grand mean	−31.1	<0.001
2	−0.1992	0.0357	$\Theta_{2}^{(1)}$: zygosity (DZ)	−5.58	<0.001
3	0.3883	0.0495	$\Theta_{12}^{(23)}$: sex (M) · country (F)	7.85	<0.001
4	0.2319	0.0457	$\Theta_{21}^{(23)}$: sex (F) · country (S)	5.08	<0.001
5	−0.2633	0.0538	$\Theta_{22}^{(23)}$: sex (F) · country (F)	4.89	<0.001

level. Thus, we may proceed to test two-way interaction hypotheses, of which there are three kinds in this case. For the hypothesis of no interaction between zygosity and country, the test statistic yielded a P value of 0.36. Thus, this hypothesis could not be rejected. Similarly, a hypothesis on no interaction between sex and zygosity could not be rejected ($P = 0.13$). When the third hypothesis on two-way interactions is considered, the hypothesis is emphatically rejected ($P < 0.001$). It may be concluded that an interaction between sex and country effects on concordance ratios for cigarette smoking exists. Because one second-order no-interaction hypothesis has been rejected, the hypotheses on main effects must be considered in light of this result. Since the hypotheses of interactions of zygosity with either sex or country were not rejected, the hypothesis of a zygosity main effect may be tested. This null hypothesis is rejected at the $P < 0.001$ level and it may be concluded that there exists a true difference in concordance ratios between MZ and DZ pairs. Since a country-sex interaction effect could be observed, the testing of the main effects for sex or country is not valid as the testing of lower-order effects presupposes that higher-order interaction hypothesis have not been rejected.

MODEL FITTING

In the choice of an optimal model, it is necessary to omit all unnecessary interaction terms, since a substantial loss of power might result from fitting them [1]. For the last model not to be rejected, the estimates and their standard errors were produced. The estimates were then statistically tested (Table 4).

The estimates were all significantly different from zero. Using these estimates the expected number of cases was computed (Table 5).

The expected numbers were very close to the observed ones. The correlations between estimated coefficients from Table 4 is shown in Table 6. The estimated coefficient for zygosity had little correlation to the other estimates, while large correlations could be observed between the sex and country interaction term estimates.

The analysis indicates that zygosity has a significant effect on concordance ratios for cigarette smoking in this study, and this effect is independent of the other study variables as seen by nonsignificant interaction terms and a lack of correlation to other coefficient estimates. The model yielded estimates for expected values of concordant pairs that were very close to those observed. The largest difference was observed for Finnish female pairs. A significant sex and country interaction was observed, indicating that the concordance ratio in relation to sex was different in the two countries. This also meant that the effect of sex and country could not be assessed independently.

TABLE 5. Observed and Expected Number of Concordant Pairs Under the Logit Model for Smoking Concordance

Variables			Observed pairs	Expected pairs
Sweden	Male	MZ	452	445.8
		DZ	600	606.2
	Female	MZ	641	637.6
		DZ	839	842.4
Finland	Male	MZ	373	357.8
		DZ	727	742.2
	Female	MZ	232	256.8
		DZ	469	444.2

TABLE 6. Correlations of Estimates (Numbered as in Table 4) for the Smoking Concordance Logit Model

1	1.000				
2	−0.518	1.000			
3	−0.559	−0.073	1.000		
4	−0.649	0.004	0.527	1.000	
5	−0.530	−0.037	0.450	0.485	1.000
	1	2	3	4	5

DISCUSSION

In this study a method of analysis was presented for analyzing the significance of familial and genetic factors in twin studies of quantitative traits. Previously, in studies where the population prevalence of the trait is known, familiarity has been assessed by comparing the observed concordance ratio for DZ pairs to that expected on the basis of independent prevalence in individuals. Genetic factors have been assessed by comparing the MZ ratio to the DZ ratio. This assumes that the common environment of MZ and DZ twin pairs is similar. Such a method of analysis disregards information on the stratum effects, as well as interaction effects. To control for other factors in genetic epidemiological studies, Kastenbaum's and Lamphiear's test [16] for no three-way interactions has been used [8]. Conclusions made solely on the basis of the above method of analysis may be misinformative since one is not justified in the statistical testing of within-strata effects, if tests for overall effects and interactions have not been carried out. Thus, a more complete and balanced view of the data is necessary.

For such a purpose, a model was applied in this paper, the logit model, that is analogous to the widely used log-linear model for contingency table analysis. This model permits testing of interaction effects before estimating the main effects of the study variables. The model may also be easily extended to four-way or even more complex tables, though the testing procedures and sequential hypothesis testing becomes increasingly demanding. Prior selection of the multiway table to be analyzed becomes necessary [9].

In the example used, data from a cross-national study of cigarette smoking among adult twin pairs in two countries was used. The multivariate assessment of genetic factors in relation to other factors was carried out by logit analysis of concordance ratios by

https://doi.org/10.1017/S0001566000006474 Published online by Cambridge University Press

analyzing three variables, zygosity, sex, and country, at the same time. Thus, the effect of sex and country on zygosity in the smoking trait could be controlled. For cigarette smoking, the present results indicate that in both countries the zygosity effect is significant, and independent of country and sex. Although not presented here, the results held for both current and ever cigarette smoking as well as for smoking over one pack a day either currently or ever. For the heavy smokers, the zygosity effect had a significant interaction with sex [15]. A significant zygosity effect implies greater concordance for the trait among MZ than DZ pairs and is due to the identical genome and probably greater common environment of MZ twin pairs.

REFERENCES

1. Aitkin MA (1978): The analysis of unbalanced cross classifications (with discussion). J Roy Stat Soc A 141.
2. Allen G, Hrubec Z (1979): Twin concordance: A more general model. Acta Genet Med Gemellol 28:3–13.
3. Andersen AH (1974): Multidimensional contingency table. Scand J Stat 1:115–127.
4. Andersen EB (1980): Discrete statistical models with social science applications. Amsterdam: North-Holland Publishing Co.
5. Baker RT, Nelder JA (1978): "The GLIM System Release 3 Manual." Oxford: Numerical Algorithms Group.
6. Benedetti JK, Brown MB (1978): Strategies for the selection of log-linear models. Biometrics 34:680–686.
7. Birch MW (1963): Maximum likelihood in three-way contingency tables. J Royal Stat Soc B25:220–233.
8. Cederlöf R, Friberg L, Lundman T (1977): The interactions of smoking, environment, and heredity and their implications for disease etiology. Acta Med Scand Suppl XXX:612.
9. Freeman DH, Jekel JF (1980): Table selection and log-linear models. J Chronic Dis 33:513–524.
10. Goodman LA (1970): The multivariate analysis of qualitative data: Interactions among multiple classifications. J Am Stat Assoc 65:226–255.
11. Goodman LA (1971): The analysis of multidimensional contingency tables: Stepwise procedures and direct estimation methods for building models for multiple classifications. Technometrics 13:33–61.
12. Haberman SJ (1972): Log-linear fit for contingency tables: Statistical Algorithm. Appl Stat 21:218–225.
13. Haberman SJ (1973): Printing multidimensional tables statistical algorithm. Appl Stat 22:118–126.
14. Haberman SJ (1974): The Analysis of Frequency Data. Chicago: The University of Chicago Press.
15. Kaprio J, Hammar N, Koskenvuo M, Floderus-Myrhed B, Langinvainio H, Sarna S (1981): Cigarette smoking and alcohol use in Finland and Sweden: A cross-national twin study (submitted for publication).
16. Kastenbaum MA, Lamphiear DE (1959): Calculation of chi-square to test the no three-factor interaction hypothesis. Biometrics 15:107–115.
17. Nelder J, Wedderburn RWN (1972): Generalised linear models. J Roy Stat Soc A 135:370–384.
18. Smith C (1974): Concordance in twins: Methods and interpretation. Am J Hum Genet 26:454–466.

Correspondence: Dr. J. Kaprio, Department of Public Health Science, Haartmaninkatu 3, SF-00290 Helsinki 29, Finland.

https://doi.org/10.1017/S0001566000006474 Published online by Cambridge University Press