Acta Genet Med Gemellol 39:173-180 (1990)
©1990 by The Mendel Institute, Rome

Sixth International Congress
on Twin Studies

Incorporation of Twins in the Regressive Logistic Model for Pedigree Disease Data

J.L. Hopper[1], J.B. Carlin[2], G.T. Macaskill[1], P.L. Derrick[1], L.B. Flander[3], G.G. Giles[4]

[1] *Faculty of Medicine Epidemiology Unit, and* [2] *Department of Community Medicine, The University of Melbourne;* [3] *Department of Social and Preventive Medicine, Monash University;* [4] *Anti-Cancer Council of Victoria, Australia*

Abstract. Segregation and twin disease concordance analyses have assumed a theoretical underlying liability following a multivariate normal distribution. For reasons of computation, of incorporation of measured explanatory variables, and of testing of fit and assumptions, newer analytical methods are being developed. The regressive logistic model (RLM) relies on expressing the pedigree likelihood as a product of conditional probabilities, one for each individual. In addition to logistic regression modelling of measured epidemiological variables on disease prevalence, there is modelling of vertical transmission, of transmission of unmeasured genotypes and of sibship environment. This paper discusses methods for the analysis of binary traits in twins and in pedigrees. Some extensions to the RLM for pedigrees which include twins are proposed. These enable exploration of twin concordance in the context of the twins' common parenthood, the sibship similarities within the family, and the twins' similarity in age, sex, genes and environment.

Key words: Binary disease data, Family data, Logistic regression, Log-linear models, Random effects, Twin data

INTRODUCTION

Segregation has been defined [6] as "the statistical methodology used to determine from family data the mode of inheritance of a particular phenotype, especially with a view to elucidating single gene effects". Following the seminal paper by Elston and Stewart [5], analysis of binary disease status data (affected, non-affected) on

https://doi.org/10.1017/S0001566000005407 Published online by Cambridge University Press

individuals in pedigrees with the aim of detecting Mendelian ratios has been based on a specific assumption. This proposes that there exists a normally distributed random variable called liability, and that an individual is affected if and only if his or her liability is greater than some parameter called the threshold. Statistical inference is based on maximum likelihood theory, and requires calculation of the likelihood of a set of observed pedigree data under an assumed model. This likelihood is expressed in terms of the multivariate distribution of the theoretical liability, and of parameters representing various modes of inheritance and factors relevant to expression of the disease, such as age, sex, and age of onset.

Statistical analyses of disease concordance in twin pairs, and in twin families, have also relied on a theoretical underlying liability, usually assumed to have a multivariate normal distribution [21]. Kramer and Corey [16] proposed a model for the analysis of twin kinship affection data based on a logistic function of liability.

It has been found that the computation time of these multivariate liability models increases rapidly as the size of pedigree increases, with severe practical limitations. These models are also limited in their ability to incorporate measured explanatory variables realistically and efficiently, and in their ability to test the fit of models and the adequacy of underlying assumptions. Consequently other approaches have been developed for the analysis of human pedigree data [eg, 9], and more generally of correlated binary data when each binary observation may have its own covariates, using generalised estimating equations [eg, 19,20]. The latter is a partially non-parametric approach, designed to deal with large "blocks". It focusses on modelling the marginal expectation (or probability of affection), treating correlations within blocks/pedigrees as nuisance parameters. It is therefore not appropriate when the correlations themselves are of scientific interest, as in pedigree analysis, because inference between different correlational representations is not possible.

The regressive logistic model (RLM) [4] relies on expressing the pedigree likelihood as a product of conditional probabilities, one for each individual. It applies logistic regression modelling to measured epidemiological variables on disease prevalence, while concurrrently allowing modelling of vertical transmission, of transmission of unmeasured genotypes and of sibship environment.

In this paper, the analysis of binary pedigree data by models which do not make the liability assumption will be examined. In particular, the RLM will be described, and some ideas applicable to families which include twins will be presented. Development of these approaches has been motivated in part by a 1968 population-based study of asthma symptoms in over 8,000 children born in 1961 attending school in Tasmania, and in their parents and siblings.

STATISTICAL MODELS

Consider a group of n related individuals. Let Z_i be the disease status (1 = affected, else 0) of individual i , $i = 1, \ldots, n$. Suppose there is a set of m measured covariates

https://doi.org/10.1017/S0001566000005407 Published online by Cambridge University Press

$\vec{X}_i = (X_{i1}, \ldots, X_{im})'$, with the matrix $\mathbf{X} = (\vec{X}_i, \ldots, \vec{X}_n)$. For convenience, the dependence of

$$\pi_i = P(Z_i = 1|\vec{X}_i) \tag{1}$$

on these, and on other explanatory variables, is modelled by a linear logistic function, ie, for any probability π, $\operatorname{logit}\pi = \log[\pi/(1-\pi)]$ is a linear function of the variables. Thus in general

$$\operatorname{logit} P(Z_i = 1|\vec{X}_i) = \gamma_0 + \gamma_1 X_{i1} + \cdots + \gamma_m X_{im}\,. \tag{2}$$

Following [8] let ρ_{ij} denote the correlation between Z_i and Z_j for any i and j (assumed to be independent of $\vec{X}_i$ and $\vec{X}_j$); ie

$$\rho_{ij} = \frac{E(Z_i Z_j) - \pi_i\pi_j}{\pi_i\pi_j(1-\pi_i)(1-\pi_j)}\,. \tag{3}$$

For a pair of individuals, each of the $2 \times 2 = 4$ probabilities covering all possible combinations of outcomes can now be written as

$$\begin{aligned} &P(Z_1, Z_2|\mathbf{X}) = \\ &= \left\{1 + \rho_{12}[\pi_1(1-\pi_1)\pi_2(1-\pi_2)]^{-1/2}(Z_1-\pi_1)(Z_2-\pi_2)\right\}\prod_{i=1}^{2}\pi_i^{Z_i}(1-\pi_i)^{1-Z_i}\,. \end{aligned} \tag{4}$$

The requirement that the expression in (4) be non-negative places severe restrictions on the range of possible values for the ρ_{ij}, for given π_i and π_j [see 11,20].

For groups of size $n = 3$, one can similarly write

$$\begin{aligned} P(Z_1, Z_2, Z_3|\mathbf{X}) = &\prod_{i=1}^{3}\pi_i^{Z_i}(1-\pi_i)^{1-Z_i}\cdot \\ &\cdot\Big\{1 + \sum_{i<j}\rho_{ij}[\pi_i(1-\pi_i)\pi_j(1-\pi_j)]^{-1/2}(Z_i-\pi_i)(Z_j-\pi_j) \\ &+ \rho_{123}[\pi_1(1-\pi_1)\pi_2(1-\pi_2)\pi_3(1-\pi_3)]^{-1/2}(Z_1-\pi_1)(Z_2-\pi_2)(Z_3-\pi_3)\Big\}\,, \end{aligned} \tag{5}$$

where

$$\rho_{123} = E[(Z_1-\pi_1)(Z_2-\pi_2)(Z_3-\pi_3)][\pi_1(1-\pi_1)\pi_2(1-\pi_2)\pi_3(1-\pi_3)]^{-1/2}\,.$$

Although certain choices of ρ_{123} require only specification of the πs and ρs and avoid restrictions [eg, 20], generalization of this approach to higher values of n is cumbersome and computationally demanding [2].

https://doi.org/10.1017/S0001566000005407 Published online by Cambridge University Press

A Log-Linear Model

A log-linear model (LLM) for binary pedigree data [10,11] allows specification of the πs in terms of measured explanatory variables, for example as a linear logistic function as in (1), and estimates the correlations ρ_{ij} under certain restrictions borrowed from log-linear modelling [7]. In the special case of pedigrees of regular size and structure, it is equivalent to fitting a log-linear model with no second or higher order interactions, and in theory this assumption can be tested by modelling higher order interactions. It has been applied to data sets consisting of pedigrees of varying sizes of up to ten individuals [10,12]. Like the multivariate normal model for continuous pedigree traits [9,17], the LLM makes assumptions about the structure of data within pedigrees which allow information to be pooled across pedigrees of arbitrary size and structure. Both are defined in terms of 'mean' components and of 'associations' between pairs of individuals, and can incorporate measured environmental and genetic variables.

The LLM has been used to analyse a pair of binary traits (having ever had asthma/hayfever) measured in almost 3,000 Australian twin pairs [15]. A higher cross-correlation among identical (MZ) pairs, compared to fraternal (DZ) pairs, between having asthma in one twin and having hayfever in the other twin, was shown to be explained by higher MZ correlations both in asthma and in hayfever. That is, there was putative evidence of genetic factors both for asthma and hayfever, and that a component of these factors was common to both allergies.

The LLM is in essence a descriptive model, and can make only indirect inference on genetic and environmental effects by reference to estimated correlations between different categories of relatives (eg, MZ vs same-sex DZ twins), as in [15]. This is in contrast to classic biometric modelling based on partitioning of the variance according to genetic and environmental sources of error.

The Regressive Logistic Model

The regressive logistic model (RLM) extends simple Markovian structures for dependence, introduced in regressive models for continuous traits [3], to binary traits through use of the logistic function [4]. The RLM allows flexible modelling of vertical transmission, and can incorporate explanatory variables and major gene effects for segregation and linkage analyses.

A limited modelling of horizontal transmission has been proposed by specifying an appropriate sequence among the siblings, and thereby invoking some kind of order in the pedigree. This has been achieved by assuming that shared sibling environment is determined statistically either by parents and older siblings, or by siblings closer in birth order. However, if there are twins in the sibship (ie, siblings of exactly the same age) this approach will break down.

For illustrative purposes, consider a nuclear family of size n consisting of at least one parent and s siblings. We shall allow for the possibility that some of these siblings may be twinned. Let $Z_1, \ldots, Z_s$ correspond to the siblings, and Z_M and

https://doi.org/10.1017/S0001566000005407 Published online by Cambridge University Press

Z_F to the mother and father respectively. The RLM breaks down the likelihood of the family, in the first instance by generation, separating parents from offspring. That is,

$$(6) \qquad P(\vec{Z}|\mathbf{X}) = P(Z_M, Z_F|\mathbf{X})P(Z_1, \ldots, Z_s|Z_M, Z_F, \mathbf{X})\,.$$

The Class A RLM assumes that the affection status of an individual depends only on that of its parents and its own measured covariates, and that the parents are independent. That is,

$$(7) \qquad P(\vec{Z}|\mathbf{X}) = P(Z_F|\vec{X}_M)P(Z_F|\vec{X}_F)\prod_{i=1}^{s} P(Z_i|Z_M, Z_F, \vec{X}_i)\,,$$

where i goes from 1 to s. The dependence of Z_i on Z_M, Z_F, and X_i is expressed in terms of logistic modelling by writing, for example,

$$(8) \qquad \operatorname{logit} P(Z_i|Z_M, Z_F, \vec{X}_i) = \alpha + \gamma_M Z_M + \gamma_F Z_F + \gamma_1 X_{i1} + \cdots + \gamma_m X_{im}\,.$$

For the parents, $j = M, F$, we may write

$$(9) \qquad \operatorname{logit} P(Z_j|\vec{X}_j) = \beta + \gamma_1 X_{j1} + \cdots + \gamma_m X_{jm}\,.$$

This model is expressed in terms of the following parameters: α, the baseline for siblings and β, the baseline for parents, the regression coefficients γ_M, γ_F, and γ_i representing the effects on log odds of risk due to an affected mother, to an affected father and to each of the measured covariates, respectively. RLMs can incorporate a major unmeasured gene [4], a common sibling environment [13,14], and several ascertainment corrections can be easily invoked [14].

Incorporation of Twins in the RLM

A method for accommodating a twin pair among the siblings (for simplicity denoted by the first two siblings) is to write the joint distribution of the twin pair, conditional on their parents, expressing any similarity after taking into consideration the parental status (which is identical for the twins) by a correlation parameter. That is,

$$(10) \qquad P(\vec{Z}|\mathbf{X}) =$$

$$= P(Z_M|\vec{X}_M)P(Z_F|\vec{X}_F)P(Z_1, Z_2|Z_M, Z_F, \vec{X}_1, \vec{X}_2)\prod_{i=3}^{s} P(Z_i|Z_M, Z_F, \vec{X}_i)\,,$$

where the joint distribution of Z_1 and Z_2 is given by (4) with π expressed by (8). Information on the vertical transmission effects are derived from consideration of

https://doi.org/10.1017/S0001566000005407 Published online by Cambridge University Press

the observed affection status of parents and all their siblings (twins included), and this is used to derive the estimated affection probability π of each individual. The correlation between twin pairs is then concurrently estimated, taking any parental effects into consideration. Therefore this model allows testing of the hypotheses that some or all of any observed concordance in affection status of siblings and of twins can be explained by common parentage, and that (categories of) twins are more highly correlated than siblings. Note that equation (10) can be extended by: (a) modelling the pair of parents in the same way as a pair of twins, and estimating a spouse concordance, and (b) proposing a LLM for the joint distribution of siblings, and estimating a common sibling concordance.

Incorporation of Random Effects in the RLM

An alternative approach, in the spirit of classical biometric modelling, is to introduce random effects to model associations between relatives due to common familial factors. Thus, for a sibship which includes a twin pair, we may rewrite (8) as

$$(11) \qquad \operatorname{logit} P(Z_i | Z_M, Z_F, \overrightarrow{X}_i, \delta, \eta) = \alpha + \gamma_M Z_M + \gamma_F Z_F + \overrightarrow{\gamma}' \overrightarrow{X}_i + \eta T_i + \delta \,,$$

where $T_i = 1$ if sibling i is one of a twin pair, and 0 otherwise, and η and δ are independent normally distributed random effects with mean 0 and variances σ_T^2 and σ_E^2 respectively. That is, associated with each sibship there is a random contribution δ to the logit which induces a correlation between siblings and increases the variance of the distribution of the number of cases in sibships. Similarly associated with the twin pair there is a further random contribution η to the logit which induces an extra correlation between twins. The likelihood of a family is now obtained by modifying (7) to include an integration of the conditional probabilities of affection given the value(s) of the random effect(s) over a univariate, or in the case of twins a bivariate, normal distribution. In the simpler case of no twin effects, the contribution to this marginal likelihood from a sibship is

$$(12) \quad P(\overrightarrow{Z} | \mathbf{X}) = P(Z_M | \overrightarrow{X}_M) P(Z_F | \overrightarrow{X}_F) \int_{-\infty}^{\infty} \prod_{i=1}^{s} P(Z_i | Z_M, Z_F, \overrightarrow{X}_i, \delta) \phi(\delta, \sigma_E^2) \, d\delta \,,$$

where $\phi(\delta, \sigma_E^2)$ is the probability density function of a normal distribution with mean 0 and variance σ_E^2. When a twin effect is incorporated, the integral in (12) becomes a double integral over two independent normal distributions:

$$(13) \quad \begin{aligned} P(\overrightarrow{Z} | \mathbf{X}) &= P(Z_M | \overrightarrow{X}_M) P(Z_F | \overrightarrow{X}_F) \cdot \\ &\cdot \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \prod_{i=1}^{s} P(Z_i | Z_M, Z_F, \overrightarrow{X}_i, \delta, \eta) \phi(\delta, \sigma_E^2) \phi(\eta, \sigma_T^2) \, d\delta \, d\eta \,. \end{aligned}$$

The integration can be handled efficiently and accurately by Gauss-Hermite numerical quadrature methods, although computation time increases substantially for the double integral. Similar methods for likelihood computations in random effects logistic regression models have been used by Anderson and Aitkin [1].

For all these models, parameter estimation and statistical inference can be performed by maximum likelihood methods using an iterative numerical procedure, such as SEARCH [18]. The use of approximate confidence intervals based on the profile likelihood is recommended because of the difficulties in applying classical tests of hypotheses at the boundary of the parameter space [1].

Acknowledgments. This work was supported by the Australian National Health and Medical Research Council (NHMRC), The Asthma Foundation of Victoria and The Asthma Foundation of Tasmania, and carried out in conjunction with the Australian NHMRC Twin Registry.

REFERENCES

1. Anderson DA, Aitkin M (1985): Variance component models with binary response: interviewer variability. J R Statist Soc B 47:203-210.
2. Bahadur RR (1961): A representation of the joint distribution of responses to n dichotomous items. In H. Solomon (ed): Studies in Item Analysis and Prediction, Stanford Mathematical Studies in the Social Sciences VI. Stanford, California: Stanford University Press.
3. Bonney GE (1984): On the statistical determination of major gene mechanisms in continuous human traits: Regressive models. Am J Med Genet 18:731-749.
4. Bonney GE (1986): Regressive logistic models for familial disease and other binary traits. Biometrics 42:611-625.
5. Elston RC, Stewart J (1971). A general model for the genetic analysis of pedigree data. Hum Hered 21:523-542.
6. Elston RC (1981). Segregation analysis. In (H. Harris, K. Hirschhorn, eds): Advances in Human Genetics, Vol. 11. New York: Plenum Press, Ch 2.
7. Fienberg SE (1977): The Analysis of Cross-Classified Categorical Data. Cambridge, MA: MIT Press.
8. Hannah MC, Hopper JL, Mathews JD (1983): Twin concordance for a binary trait. I. Statistical models illustrated with data on drinking status. Acta Genet Med Gemellol 332:127-137.
9. Hopper JL, Mathews JD (1982). Extensions to multivariate normal models for pedigree analysis. Ann Hum Genet 46:373-383.
10. Hopper JL, Hannah MC, Mathews JD (1984): Genetic Analysis Workshop II. Pedigree analysis of a binary trait without assuming a liability. Genet Epidemiol 1:183-188.
11. Hopper JL, Derrick PL (1986): A log-linear model for binary pedigree data. Genet Epidemiol Supplement 1:73-82.
12. Hopper JL, Judd FK, Derrick PL, Burrows GD (1987): A family study of panic disorder. Genet Epidemiol 4:33-41.
13. Hopper JL (1989): Modelling sibship environment in the regressive logistic model for familial disease. Genet Epidemiol 6:235-240.
14. Hopper JL, Judd FK, Derrick PL, Macaskill GT, Burrows GD (1990): A family study of panic disorder - Reanalysis using a regressive logistic model that incorporates a sibship environment. Genet Epidemiol (in press).
15. Hopper JL, Hannah MC, Macaskill GT, Mathews JD (1990): Twin concordance for a binary trait. III. A bivariate analysis of hayfever and asthma. Genet Epidemiol (in press).

https://doi.org/10.1017/S0001566000005407 Published online by Cambridge University Press

16. Kramer AA, Corey L (1986). The offspring of twins as sampling units in pedigree analysis of congenital anomalies. Acta Genet Med Gemellol 35:35-48.
17. Lange K, Westlake J, Spence MA (1976): Extensions to pedigree analysis. III. Variance components by the scoring method. Ann Hum Genet 46:373-383.
18. Lange K, Boehnke M, Weeks D (1986): Programs for Pedigree Analysis. UCLA School of Biomathematics, Los Angeles.
19. Liang KY, Zeger SL (1986): Longitudinal data analysis using generalized linear models. Biometrika 73;13-22.
20. Prentice RL (1988): Correlated binary regression with covariates specific to each binary observation. Biometrics 44:1033-1048.
21. Smith C (1974): Concordance in twins: Methods and interpretation. Am J Hum Genet 26:454-466.

Correspondence: Dr. John L. Hopper, The University of Melbourne, Faculty of Medicine Epidemiology Unit, 151 Barry Street, Carlton, Victoria 3053, Australia.

https://doi.org/10.1017/S0001566000005407 Published online by Cambridge University Press