Acta Genet Med Gemellol 36:21-27 (1987)

Innovations in the Statistical Analysis of Twin Studies

J.L. Hopper, P.L. Derrick, C.A. Clifford

Faculty of Medicine Epidemiology Unit, University of Melbourne, Carlton, Victoria, Australia

Abstract. Advances in computer technology have made possible a greater sophistication in the statistical analysis of pedigree data, however this is not necessarily manifest by fitting more comprehensive causative models. Planned twin and family studies measure numerous explanatory variables, including perhaps genetic and DNA marker information status on all pedigree members, and the cohabitation of all pairs of individuals. A statistical analysis should examine the contribution of these measured factors on individual means, and in explaining the variation and covariation between individuals, concurrently with the postulated effect of unmeasured factors such as polygenes. We present two models that meet this requirement: the Multivariate Normal Model for Pedigree Analysis for quantitative traits, and a Log-Linear Model for Binary Pedigree Data. For both models, important issues are examination of fit, detection of outlier pedigrees and outlier individuals, and critical examination of the model assumptions. Procedures for fulfilling these needs and examples of modelling are discussed.

Key words: Binary data, Log-linear models, Logistic regression, Maximum likelihood, Multivariate normality, Pedigree analysis

INTRODUCTION

In recent years the availability of fast computation has made it feasible to fit more sophisticated models in the statistical analysis of pedigree data. For example, following Elston and Stewart [3], algorithms that incorporate maximisation routines have been used to fit a variety of genetic models, using maximum likelihood theory. However, although there

This work was supported by the Australian National Health and Medical Research Council.

has been a tendency to fit more comprehensive causative models, this is not the only nor most informative way in which the increased computer power can be utilised.

Properly conducted twin and family studies measure numerous explanatory variables as a matter of course. In particular, genetic or DNA marker information on all pedigree members, and the cohabitation history of all pairs of individuals, might be collected. A proper statistical analysis should examine the contribution of measured factors, a) on mean values, and b) in explaining the (co)variation between individuals, concurrently with the postulated effect of unmeasured factors such as polygenes. For quantitative traits the multivariate normal model for pedigree analysis [11,20], and for binary traits a log-linear model for binary pedigree data [7,8], meet this requirement. For both models important issues are: examination of fit, detection of outlier pedigrees, detection of outlier individuals, the effect of departures from model assumptions, and testing of both statistical and biological model assumptions.

The classical twin method makes the assumption that the effect of shared environment is the same in monozygotic (MZ) pairs as in dizygotic (DZ) pairs. This assumption cannot be tested from twin data alone, yet it is an observation that on average MZ twins have more similar lifestyles and more similar environments. It has been argued that this could be, at least in part, a consequence of their greater genetic similarity [16]. However, notwithstanding the cause, this observation biases the twin method. If the correlation between MZ pairs, r(MZ), is greater than the correlation between DZ pairs, r(DZ), this does not "prove" a genetic hypothesis, while if r(MZ) = r(DZ) this does not necessarily "disprove" a genetic hypothesis.

As a strategy for overcoming some of the weaknesses of the classical twin method, it has been proposed that designs involving twin families or a combination of twin and family data (which provide contrasts between genetic factors and common environment factors and are therefore more informative designs than twins alone) be used. Therefore the analysis of twin studies can be viewed as a special case of pedigree analysis.

THE MULTIVARIATE NORMAL MODEL FOR PEDIGREE ANALYSIS

Consider a pedigree of size n, on which a continuous trait X with observed values $x = (x_1, x_2, ..., x_n)'$ has been measured. The expected values are $E[X] = \mu = (\mu_1, \mu_2, ..., \mu_n)$, where μ_j, (j = 1,2, ..., n) may depend on measured explanation variables, such as the age, sex, and other characteristics of individual j. It is assumed that X is distributed as a multivariate normal variate with mean μ and variance-covariance matrix Ω, where Ω depends on the relationship between members of the pedigree, and on the proposed causal model.

A linear model for the fixed effects assumes that

$$\mu_j = \beta_{0j} + \beta_1 y_{1j} + \beta_2 y_{2j} + ... + \beta_p y_{pj},$$

for individual j where y_{kj} is the value of the kth. explanatory variable. As suggested by Hopper and Mathews [11], of particular importance in genetic modelling is the case where these explanatory variables represent one or more measured genetic loci with at most several alleles [2,22].

Most causal models make the assumption that the "random" effects are independent

and additive, and as such the variance-covariance matrix can be decomposed into linear components; ie, $\Omega = \Sigma\ \theta_i \Omega_i$. The parameters θ_i represent the random effects, and could themselves be functions of other measured variables, such as the ages of pedigree members [23,25]. To model unmeasured additive genetic effects, $\Omega_i = 2\phi$ where ϕ is the kinship matrix [20].

For a measured genetic marker that is highly polymorphic, eg, HLA, $\Omega_i = (\psi_{jk})$ where ψ_{jk} is 1 if individuals j and k share both haplotypes, or 0.5 if j and k share one haplotype [11,12].

For a cohabitation effect, many parameterisations are possible, [eg, 24]. One possible parameterisation involves letting $\Omega_i = (\gamma_{jk})$, where if time t is measured from when individuals j and k first begin cohabiting and t_0 is the time at which they may have ceased to cohabitate, then γ_{jk} is $1 - e^{-\lambda t}$ if $t < t_0$, otherwise $(1 - e^{-\lambda t_0})e^{-\nu(t-t_0)}$ [11]. This parameterisation allows a large range of possible shapes; [6:Fig.1]. Furthermore, a stochastic mechanism has been proposed that causes covariances to converge or diverge exponentially fast as relatives cohabit or lead separate lives [2,19].

Considerable cohabitation effects have been observed in analyses of family data using this parameterisation. For example, a large difference in correlation between cohabiting and non-cohabiting siblings was evident for blood lead levels [12], while a disaggregating effect attributed to cohabitation was detected for Cattell's personality factor A (sizia versus affectia) for mother-offspring pairs [6]. Large cohabitation effects for MZ, and to a lesser extent DZ twins were observed in anxiety and depression symptom scores, and in alcohol consumption of drinking twins, in a study of UK twin families [4].

It is possible also to derive the Ω_i by reference to a path diagram [26].

Parameters are estimated by maximimum likelihood methods. The log likelihood of a pedigree i is to a constant

$$\log L_i = -1/2\ [\log|\Omega| + (x - \mu)'\Omega^{-1}(x - \mu)].$$

If it is the case that pedigrees are independently sampled, the sample log likelihood is $\ell = \Sigma \log L_i$. It is a function $\ell = \ell(\beta, \theta, \lambda, \nu)$ of parameters representing the effects of measured explanatory variables, the effects of unmeasured factors that influence covariation, and aspects of the effects of cohabitation. Maximization of ℓ with respect to a parameter space is achieved by an iterative method; for example by direct search using MAXLIK [15], or by Quasi-Newton methods using SEARCH [19].

Is is important to consider the fit to the multivariate normal distribution, because (i) skewness influences estimates of the mean, and (ii) kurtosis influences estimates of standard errors [14]. For example, we conducted a simulation study for samples of 20 pedigrees of size 5 according to the simple model $X_{ij} = A_i + E_{ij}$, (i = 1, ..., 20; j = 1, ..., 5) where A and E are independent Student's t variates on 6 degrees of freedom with variances σ_a^2 and σ_e^2 respectively, $\sigma^2 = \sigma_a^2 + \sigma_e^2$. It was found that as the within-pedigree correlation, $\rho = \sigma_a^2/\sigma^2$, increased from 0 to 1/5 to 1/3 the degree of underestimation in estimates of the standard error for ρ increased from 2.7% to 14.4% to 19.2%. However, the marginal kurtosis decreased from 3.0 to 2.0 to 1.67. Therefore, kurtosis in the component induces underestimation in standard error estimates, but examination of marginal kurtosis alone may not reveal this even when there is small within-pedigree correlation.

It is of interest, however, to notice that the estimates of ρ itself are reasonably robust

to skewness [26], and to kurtosis. In the simulation study above, the mean of the estimates for ρ were 0.00, 0.19 and 0.32 respectively.

Hopper and Mathews [11] introduced some proposals for assessing fit. For each pedigree i, decompose $\Omega_i = B_i \Lambda_i B_i'$, Λ_i diagonal. Let $z_i = \Lambda_i^{-1/2} B_i'(x_i - \mu_i)$, and replace Λ_i, B_i and μ_i by their respective maximum likelihood estimates. A test of the multivariate normal assumption is provided by examining a plot of the ordered (z_{ij}) against the expected normal order statistics [11]; in particular the correlation provides an easily calculated omnibus test statistic [14]. A test for outlier pedigrees is based on examination of $Q_i = z_i' z_i$ [11,18]. A test for multivariate kurtosis can be based on ΣQ_i^2 [14,21].

A LOG-LINEAR MODEL FOR BINARY PEDIGREE DATA

Consider binary pedigree data. For individual j let $Z_j = 1$ if disease is present, else 0. The disease probability $\pi_j = P(Z_j = 1)$, is either the prevalence or the cumulative risk, depending on sampling considerations. It has been the practice to analyse binary pedigree data by liability models [27,29]. Although this approach has been applied extensively there are several drawbacks: 1) one can never test the assumption of multivariate normality of liability, yet the estimation procedure is model dependent; 2) there are problems in the numerical approximations used for calculating the tails of multivariate normal distributions; and 3) it is not clear whether adjustments for measured covariates such as age of onset and sex should be made to threshold, or to the variance of liability.

Other authors [28,30] have sought to develop methods that might not have some of these difficulties. We have proposed a descriptive model for binary pedigree analysis [8] which does not make the multivariate normal liability assumption, is numerically stable, and allows for adjustments for measured covariates to be made to the disease probability. This model makes the assumption: for every pair of individuals j and k, the odds ratio

$$\psi_{jk} = P(Z_j = 1, Z_k = 1)\, P(Z_j = 0, Z_k = 0) / P(Z_j = 1, Z_k = 0)\, P(Z_j = 0, Z_k = 1)$$

is independent of any event involving neither Z_j nor Z_k. This structure is motivated by considering log-linear models with no second- or higher-order interactions. If the disease is rare, ψ approximates the relative risk of an individual being affected for the presence of an affected relative, and our assumption above is almost equivalent to relative risks being multiplicative. In theory it is possible to test this assumption by fitting higher-order interactions and for small pedigrees this is feasible [9].

From an epidemiological point of view it is convenient to express disease concordance between a pair of relatives in terms of the odds ratio, and for computational purposes the "natural" scales are log ψ and logit π. As an example of the model, analysis of a family study of panic disorder [10] showed that after adjusting the disease probability for age of onset by logistic regression, the presence of an affected first-degree relative induced about a five-fold increase in the risk of being affected, irrespective of whether the relationship was between parent and offspring or between sibling pairs.

Although the model is descriptive, genetic/environmental interpretations can be introduced. In particular, the model can accomodate genetic markers, either by modelling the disease probability as a function of alleles or genotypes, or by allowing concordance

between pairs of individuals to be a function of the number of alleles or haplotypes shared [8]. Cohabitation history can also be modelled in a similar way to that suggested above for the multivariate normal modelling of continuous traits [5].

The model fit can be assessed to some extent by comparing the number of cases in each pedigree with the number expected under the fitted model, and as a consequence atypical pedigrees can be identified. As mentioned above it is possible to test for higher-order interactions, but for large pedigrees this involves introducing an additional large number of parameters, with a consequent increase in computational time. Tests have shown that computation time increases rapidly with the size of pedigree, n, at a rate greater than $(2.5)^n$. For our current implementation it is necessary to work with pedigrees of 10 or less individuals, unless simplifying assumptions [eg, 8] can be used to decompose larger pedigrees.

DISCUSSION

In the near future a large amount of genetic marker information will be routinely collected as part of pedigree studies. In practice it is a matter of form for researchers to collect data on variables that are known or hypothesised to be important. Therefore, it is logical that these should be examined in an analysis, concurrently with studying familial aggregation. Both the methods discussed above, for continuous and for binary pedigree data, can be applied to samples of pedigrees of arbitrary size and structure, and allow for measured factors to be incorporated in the analysis concurrently with the postulated effect of unmeasured factors such as polygenes.

In several examples significant differences between the correlation between cohabiting and non-cohabiting pairs of relatives have been found, the more so for twin pairs. These effects related to cohabitation have been shown to have a considerable influence on the genetic/environmental interpretation of trait variation [4]. It is possible that in particular cases there would be other measured factors which influence trait covariation, and the methods of analysis proposed would allow these to be investigated.

Although there has been a tendency in biometrical modelling for researchers to develop more comprehensive causative models (conveniently represented by path analysis diagrams) they have invariably invoked hypothetical, although intuitively realistic, unmeasured factors. These models can be analysed by the multivariate normal model for pedigree analysis [25] and it is suggested that attention should given to complementing these models by including measured factors in model specification and in analysis.

Acknowledgements. The authors wish to acknowledge the contributions of Prof. J.D. Mathews and Mr. M.C. Hannah in development of the methodology.

REFERENCES

1. Boerwinkle E, Chakraborty R, Sing CF (1986): The use of measured genotype information in the

analysis of quantitative phenotypes in man. I. Models and analytical methods. Ann Hum Genet 50:181-194.
2. Eaves LJ, Long J, Heath AC (1986): A theory of developmental change in quantitative phenotypes applied to cognitive development. Behav Genet 16:143-162.
3. Elston RC, Stewart J (1971): A genetic model for the genetic analysis of pedigree data. Hum Hered 21:523-542.
4. Clifford CA, Hopper JL, Fulker DW, Murray RM (1984): A genetic and environmental analysis of a twin family study of alcohol use, anxiety, and depression. Genet Epidemiol 1:63-79.
5. 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 32:127-137.
6. Hopper JL, Culross PR (1983): Covariation between family members as a function of cohabitation history. Behav Genet 13:459-471.
7. Hopper JL, Derrick PL (1986): A log-linear model for binary pedigree data. In: Genetic Analysis Workshop IV. New York: Alan R Liss.
8. Hopper JL, Hannah MC, Mathews JD (1984): Genetic Analysis Workshop II: Pedigree analysis of a binary trait without assuming underlying liability. Genet Epidemiol 1:183-188.
9. Hopper JL, Hannah MC, Mathews JD (1986): Twin concordance for a binary trait. III. A bivariate analysis of hayfever and asthma. Genet Epidemiol (submitted).
10. Hopper JL, Judd FK, Derrick PL, Burrows GD (1986): A family study of panic disorder. Genet Epidemiol 4:33-41.
11. Hopper JL, Mathews JD (1982): Extensions to multivariate normal models for pedigree analysis. Ann Hum Genet 46:373-383.
12. Hopper JL, Mathews JD (1983): Extensions to multivariate normal models for pedigree analysis. II. Modeling the effect of shared environment in blood lead levels. Am J Epidemiol 117:344-355.
13. Hopper JL, Tait BD, Propert DN, Mathews JD (1982): Genetic analysis of systolic blood pressure in Melbourne families. Clin Exp Pharmacol Physiol 9:247-252.
14. Hopper JL, Speed TP (1986): Testing of assumptions in the simplest variance component model. Biometriks (submitted).
15. Kaplan EB, Elston RC (1972): A subroutine package for maximum likelihood estimation (MAXLIK). Institute of Statistics Mimeo Series No. 823 (revised March 1978).
16. Kendler KS, Heath A, Martin NG, Eaves LJ (1986): Symptoms of anxiety and depression in a volunteer twin population. Arch Gen Psychiatry 43:213-221.
17. Lange K (1986): Cohabitation, convergence, and environmental covariances. Am J Med Genet 24:483-491.
18. Lange K, Boehnke M (1983): Extensions to pedigree analysis. IV. Covariance components models for multivariate traits. Am J Med Genet 14:513-524.
19. Lange K, Boehnke M, Weeks D (1986): Programs for Pedigree Analysis. Los Angeles: Department of Biomathematics, UCLA School of Medicine.
20. Lange K, Westlake J, Spence MA (1976): Extensions to pedigree analysis. III. Variance components by the scoring method. Ann Hum Genet 39:485-491.
21. Mardia KV (1974): Applications of some measures of multivariate skewness and kurtosis in testing normality and robustness studies. Sankhya Ser B 36:115-128.
22. Martin NG, Clark P, Ofulue AF, Eaves LJ, Corey LA, Nance WE (1986): Does the Pi polymorphism alone control alpha-1-antitripsin expression? Am J Hum Genet 40:267-277.
23. Mathews JD, Hopper JL (1982): Age-dependent means and variances of quantitative traits studied over pedigrees. In: Proceedings of the Workshop on Methods for the Analysis of Family Data. University of Melbourne, pp 129-131.
24. Moll PP, Powsner R, Sing CF (1979): Analysis of genetic and environmental sources of variation in serum cholesterol in Tecumseh, Michigan. V. Variance components estimated from pedigrees. Ann Hum Genet 42:343-354.
25. Province MA, Rao DC (1985): Path analysis of family resemblance with temporal trends: Applications to height, weight and Quetelet index in Northeastern Brazil. Am Hum Genet 37:178-192.
26. Rao DC, McGue M, Wette R, Glueck CJ (1984): Path analysis in genetic epidemiology. In Chakravarti (ed): Human Population Genetics: The Pittsburgh Symposium. Stroudsburg, PA: Hutchinson Ross.

27. Rice J, Reich T (1985): Familial analysis of qualitative traits under multifactorial inheritance. Genet Epidemiol 2:301-315.
28. Self SG, Prentice RL (1986): Incorporating random effects into multivariate relative risks regression models. In Moolgavakar SH, Prentice RL (eds): Modern Statistical Methods in Chronic Diseases Epidemiology. New York: Wiley, pp. 167-177.
29. Smith C (1970): Heritability of liability and concordance in monozygous twins. Ann Hum Genet 34:85-91.
30. Thomas DC, Langholz B, Mack T, Deapen D, Floderus-Myrhed B (1986): Bivariate lifetable models for analysis of gene-environment interaction in twins. Paper presented at the Fifth Int Congr on Twin Studies, Amsterdam.

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