Acta Genet Med Gemellol 32:127–137 (1983)
The Mendel Institute/Alan R. Liss, Inc.

Received 14 March 1983

Twin Concordance for a Binary Trait.
I. Statistical Models Illustrated With Data on Drinking Status

Murray C. Hannah, John L. Hopper, and John D. Mathews

Department of Medicine, University of Melbourne, Royal Melbourne Hospital, Victoria, Australia

A flexible method based on maximum likelihood theory is introduced for the analysis of binary response data in twins. The method allows for explanatory variables such as age and sex, is free of the untestable distributional assumption of bivariate normality of liability, and makes more efficient use of the data available. The method is illustrated with preliminary data on drinking status in adult twins. Although there is some bias in the ascertainment of male dizygous twins, the results suggest that monozygous twins are more concordant than dizygous twins for drinking status.

Key words: Twin concordance, Binary trait, Correlation coefficient, Proband concordance rate, Age adjustment, Logistic regression, Transformation of scale, Likelihood, Alcohol

INTRODUCTION

Suppose we are interested in a binary trait, typically a disease or condition which may be diagnosed as present or absent. A useful way to gather information regarding the influence of genes on the occurrence of a trait is to compare the concordance rates for monozygotic (MZ) and dizygotic (DZ) twins. Any differences occurring between the individuals of an MZ twin pair must be attributable to the environment alone, since their genetic composition is the same. Furthermore, if we are prepared to assume that the environmental variability within pairs is the same for both types of twins, then the degree to which the within-pair MZ correlation exceeds the within-pair DZ correlation should reflect the extent to which genetic variability effects the occurrence of the trait.

If the trait is rare, to overcome problems associated with the ascertainment of affected individuals it is usual to base inference on the proband concordance rate defined as the proportion of co-twins with the trait for affected individuals independently ascertained. This definition is simple in practice and leads to an estimate of concordance which is independent of the ascertainment probability [11].

Often the probability of trait occurrence depends upon other explanatory variables, such as age, which are perfectly matched within pairs, and which cannot properly be regarded as being either "genetic" or "environmental" causes of variation. Such confounding variables will tend to artificially inflate the concordance rates for both MZ and

DZ twins. For example, if the trait or disease incidence increases with age, the concordance rate will tend to be higher among older twins simply due to the effect of age on the probability of occurrence, and not because of any effect of age on the intrinsic correlation between twins. Therefore, as argued by Smith [11] and others, it is important to distinguish between the concordance rate and the intrinsic correlation between twins. In this paper we describe a flexible and elegant parameterisation which emphasises this distinction, which is efficient in its use of the data, and which can be generalised to allow for the effects of either qualitative or quantitative confounding variables on the concordance rate. We illustrate the method with twin data on drinking status.

MATERIALS AND METHODS

Statistical Model

Consider a binary trait measured on a single pair of twins. Let X_k be 1 if the trait is present, and 0 if the is absent in twin k, with Π_k the probability of occurrence of the trait twin k, k=1,2. The probabilities, P_{ij}, of all four possible outcomes may be summarised in a 2×2 table (Table 1).

Each P_{ij} may be written as a function of Π_1, Π_2 and the within pair correlation coefficient ρ. To derive this function, consider

$$P_{11} = E(X_1X_2) = Cov(X_1,X_2) + E(X_1)E(X_2) = \rho\sigma_1\sigma_2 + \Pi_1\Pi_2,$$

where $\sigma_i^2 = \Pi_i(1-\Pi_i)$, i=1,2 [10]. This implies

$$P_{11} = \Pi_1\Pi_2 + \rho\{\Pi_1(1-\Pi_1)\Pi_2(1-\Pi_2)\}^{1/2}.$$

For simplicity of exposition, suppose Π is dependent upon only age and sex, so that for like-sex twin pairs $\Pi_1 = \Pi_2$, and from Table 1,

$$\left.\begin{aligned} P_{11} &= \Pi^2 + \rho\Pi(1-\Pi), \\ P_{12} &= \Pi(1-\Pi) - \rho\Pi(1-\Pi) = P_{21}, \\ P_{22} &= (1-\Pi)^2 + \rho\Pi(1-\Pi). \end{aligned}\right\} \quad (1)$$

Under this model the distinction between the proband concordance rate, P_c, and the correlation, ρ, is clearly seen if we write

$$P_c = P\{X_2=1 \mid X_1=1\} = \Pi + \rho(1-\Pi). \quad (2)$$

P_c is equivalent to the conditional probability of the second twin being affected, given that the first is affected (or vice versa). It is seen that P_c is dependent on Π, the probability of being affected, as well as on ρ, the correlation coefficient.

Having established a basic model for a single twin pair, we extend it to describe the essentials of a heterogeneous sample of N like-sex twin pairs. Logistic regression [3] may be used to model the dependence

TABLE 1. Probabilities for a Binary Trait Measured on a Twin Pair

		Twin 1		
		$X_1 = 1$	$X_1 = 0$	
Twin 2	$X_2 = 1$	P_{11}	P_{12}	Π_2
	$X_2 = 0$	P_{21}	P_{22}	$1 - \Pi_2$
		Π_1	$1 - \Pi_1$	1

X_k is 1 if the trait is present, and 0 if the trait is absent in twin k, and Π_k is the probability of occurrence of the trait in twin k, k=1,2.

https://doi.org/10.1017/S0001566000006413 Published online by Cambridge University Press

of Π on a vector of explanatory variables $\underset{\sim}{z}$. To do this we write $y = \underset{\sim}{\alpha}' \underset{\sim}{z}$, where $\underset{\sim}{\alpha}$ is a vector of real constants, and

$$\Pi = e^y / (1 + e^y). \tag{3}$$

In general, ρ may be different for each combination of sex and zygosity and, like Π, may also depend upon explanatory variables such as age (see Appendix A). However, here we will assume ρ to be dependent on sex and zygosity alone.

Thus for each twin pair the probability of the observed outcome may be written in terms of the above parameters and the observed explanatory variables. Under the (weak) assumption of independence twin pairs, the log likelihood function of the parameters given the entire sample of $(\underset{\sim}{x}, \underset{\sim}{z})$ N like-sex MZ and DZ twins is the sum of the logs of these probabilities:

$$LL(\underset{\sim}{\rho}, \underset{\sim}{\alpha}; \underset{\sim}{x}, \underset{\sim}{z}) = \sum_{n=1}^{N} \log_e P\{\underset{\sim}{x}_n; \underset{\sim}{z}_n, \underset{\sim}{\rho}, \underset{\sim}{\alpha}\}, \tag{4}$$

where $\underset{\sim}{x}_n = (x_{n1}, x_{n2})$ is the observed binary trait vector for the n^{th} twin pair, $\underset{\sim}{z}_n$ is the vector of explanatory variables for the n^{th} pair, $\underset{\sim}{\rho} = (\rho_{mmz}, \rho_{fmz}, \rho_{mdz}, \rho_{fdz})$ and $\underset{\sim}{\alpha} = (\underset{\sim}{\alpha}_m, \underset{\sim}{\alpha}_f)$, the subscripts referring to the combinations of sex and zygosity. The probabilities in (4) are calculated using the parameterisations (1) and (3). A computer routine (eg [5]) is needed to maximize the log likelihood and obtain the maximum likelihood estimate (MLE) of all parameters. LL is a regular function composed of log, exponential and polynomial functions, and inference can be drawn using results of asymptotic likelihood theory: asymptotic normality of estimators, likelihood ratio tests for model selection, and so on (see, for example, [4]). In Appendix C some MLEs and their asymptotic standard errors are derived analytically for the simplest case.

RESULTS

An Application to Data on Alcohol Consumption

To illustrate the utility of this model we have applied it to data on the drinking habits of a sample of 181 pairs of like-sex adult twins who attended a voluntary interview and medical examination as part of a study of cardiovascular risk factors. This sample was ascertained from twins living in Melbourne and registered with the Australian National Health and Medical Research Council Twin Registry. The invited sample consisted of equal numbers of male and female, and equal numbers of MZ and DZ pairs, but otherwise was drawn at random from the registry.

Information on drinking habits was obtained using a self-administered questionnaire and the responses, checked for completeness and consistency at face-to-face interview, were used to calculate alcohol consumption in grams per week. For these analyses twins have been arbitrarily categorised as nondrinkers if they consume less than 30 gm alcohol per week, and otherwise as drinkers; the binary outcome is drinking status (drinker or nondrinker).

The numbers of twin pairs in the study, by sex, zygosity and drinking status, are given in Table 2. Seventy-five pairs were invited for interview from each sex by zygosity category. However, the last line of Table 2 shows a significant sex by zygosity interaction

TABLE 2. Zygosity, Sex, and Drinking Habits of 181 Like-Sex Twin Pairs

	Male		Female		
	MZ	DZ	MZ	DZ	Total
Neither is a drinker	19	7	23	28	77
Discordant for drinking	14	16	11	15	56
Both are drinkers	19	8	11	10	48
Total	52	31	45	53	181

https://doi.org/10.1017/S0001566000006413 Published online by Cambridge University Press

in the number of pairs responding ($P \simeq 0.03$), and therefore we must be conscious of possible selection biases when interpreting the results.

A preliminary log-linear analysis, using GLIM (Generalized Linear Interactive Modeling) [1], of the proportion of drinkers, conditional upon the observed numbers in each sex by zygosity subtotal, shows a clear sex difference ($P \simeq 0.003$) but no significant zygosity effect ($P \simeq 0.8$) nor a sex by zygosity interaction ($P \simeq 0.6$).

The 362 sample individuals range in age between 18 and 70 years. Plots of the (logit) proportion of drinkers against age-group are suggestive of a quadratic relationship (Figs. 1,2).

Table 3 lists the parameters of our basic "saturated" model (Model 1), together with their MLEs and standard errors (estimated using the inverse observed information matrix; see [4,5]. For each sex a separate logistic regression model is fitted, given by $y = \kappa + \alpha t + \beta t^2$ where t is (a simple monotone function of) age in years (see Appendix B). In Table 3, Π_0 is the inverse logit of the constant term κ; $\Pi_0 = e^{\kappa}/(1+e^{\kappa})$.

Differences Between Males and Females

Table 3 indicates that there is little difference between the correlation estimates for males and females, with a possible exception for the DZ twin correlations. We test for a difference between these two correlations by using the asymptomatically standard normal variate

$$Z = \frac{\hat{\rho}_{mdz} - \hat{\rho}_{fdz}}{\{\text{Var } \hat{\rho}_{mdz} + \text{Var } \hat{\rho}_{fdz}\}^{1/2}} = -1.65,$$

which implies $P \simeq 0.10$ (two-sided).

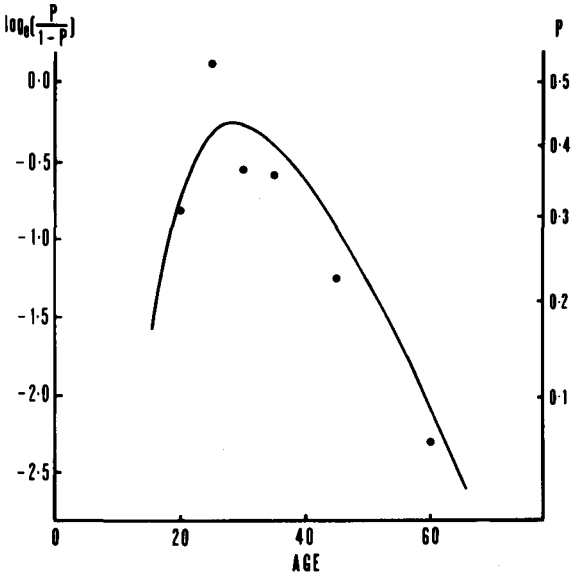

Fig. 1. Logit of P, the proportion of drinkers among males, against age. Plotted points represent the observed proportion of drinkers in age groups of at least twenty individuals, and are plotted against the mean age of the group. The continuous curve is given by the fitted logistic model for males (Model 1, Table 3) with the age transformation given in Appendix B.

https://doi.org/10.1017/S0001566000006413 Published online by Cambridge University Press

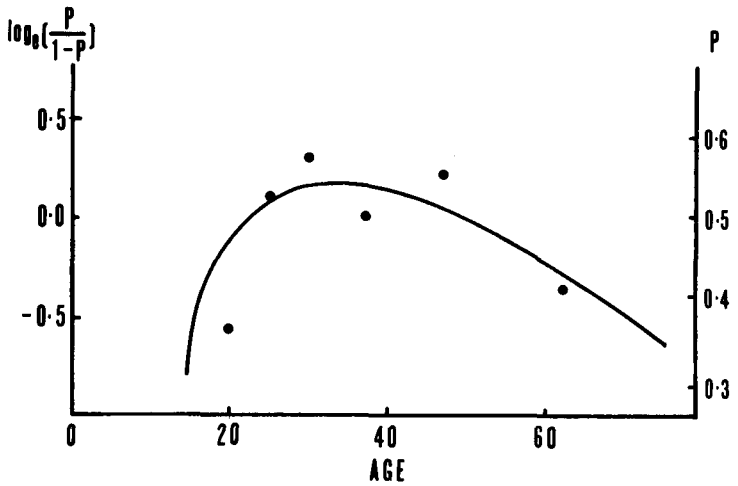

Fig. 2. Logit of P, the proportion of drinkers among females, against age. Plotted points represent the observed proportion of drinkers in age groups of at least twenty individuals, and are plotted against the mean age of the group. The continuous curve is given by the fitted logistic model for females (Model 1, Table 3) with the age transformation given in Appendix B.

TABLE 3. Parameters and Estimates for the Saturated Model, Model 1

		MLE[a]	SE[b]
Male MZ correlation	ρ_{mmz}	0.46	0.12
Female MZ correlation	ρ_{fmz}	0.42	0.15
Male DZ correlation	ρ_{mdz}	−0.05	0.18
Female DZ correlation	ρ_{fdz}	0.33	0.14
Male mean probability	Π_{mo}	0.47	0.10
Female mean probability	Π_{fo}	0.34	0.08
Male linear age adjustment	α_m	0.62	0.83
Female linear age adjustment	α_f	1.18	0.84
Male quadratic age adjustment	β_m	−0.31	0.37
Female quadratic age adjustment	β_f	−0.87	0.42

The maximized log likelihood is $LL_1 = -183.10$.
[a]Maximum likelihood estimate.
[b]Asymptotic standard error, derived from observed inverse information matrix.

There is little difference between the male and female MZ twin correlation estimates, which if pooled give a new model (Model 2) with

$$\hat{\rho}_{mz} = 0.45 \text{ (SE} = 0.09) \text{ and } LL_2 = -184.02.$$

Pooling the DZ correlations (as well as the MZ correlations) across sexes (Model 3) gives

$$\hat{\rho}_{mz} = 0.44 \text{ (SE} = 0.09), \hat{\rho}_{dz} = 0.18 \text{ (SE} = 0.11), \text{ and } LL_3 = -185.36.$$

If each of the remaining parameters, Π_o, α and β, are combined over sex (Model 4) there is a decrease in the log likelihood to $LL_4 = -190.62$. This may be used to perform

https://doi.org/10.1017/S0001566000006413 Published online by Cambridge University Press

a likelihood ratio (LR) test [3,4]: $-2(LL_4 - LL_3) = 10.54$, which as a χ^2_3 variate gives P $\leqslant 0.02$, indicating a significant difference in drinking habits between sexes.

Differences Between MZ and DZ Correlations

Although the DZ correlation estimates for males and females are suggestive of a sex difference, both are strictly less than the estimates for the MZ correlations, and there is some justification for comparing the pooled values $\hat{\rho}_{mz}$ and $\hat{\rho}_{dz}$ given under Model 3 above.

$$Z = \frac{\hat{\rho}_{mz} - \hat{\rho}_{dz}}{\{\text{Var}\,\hat{\rho}_{mz} + \text{Var}\,\hat{\rho}_{dz} - 2\text{Cov}(\hat{\rho}_{mz},\hat{\rho}_{dz})\}^{1/2}} = 1.80, \tag{5}$$

implying P $\simeq 0.036$ (one-sided).

The model with only one correlation parameter for all sex and zygosity categories (Model 5), has log likelihood $LL_5 = -186.96$. The LR test for a general difference between ρ_{mz} and ρ_{dz} is given by comparing $-2(LL_5 - LL_3) = 3.22$ with the χ^2_1 distribution, which implies P $\simeq 0.073$. Due to the nondirectionality of the LR test this is about twice the P value obtained using (5).

Using the estimates and standard errors given for Model 1 (Table 3) it is seen that, taken individually, the difference between MZ and DZ correlations is significant for males (P $\simeq 0.01$, one-sided), but not significant for females (P $\simeq 0.34$, one-sided).

The Adjustment for Age

For Model 6 the age adjustment terms are excluded from the saturated model and the estimates given in Table 4. The correlation estimates are slightly inflated due to the absence of any age adjustment.

TABLE 4. Parameters and Estimates With No Age Adjustment, Model 6

	MLE	SE	100 × SE / MLE
ρ_{mmz}	0.46	0.12	26%
ρ_{fmz}	0.47	0.14	30%
ρ_{mdz}	−0.03	0.18	—
ρ_{fdz}	0.36	0.14	39%
Π_m	0.51	0.04	8%
Π_f	0.35	0.04	11%

The maximized log likelihood is $LL_6 = -188.49$.

TABLE 5. Liability Threshold Model Estimates

	Proband concordance rate	r	$V_r^{1/2}$	100 × $V_r^{1/2}$ / r
Male MZ	0.73	0.65	0.33	51%
Female MZ	0.67	0.80	0.43	54%
Male DZ	0.50	−0.02	0.45	—
Female DZ	0.57	0.60	0.37	62%

https://doi.org/10.1017/S0001566000006413 Published online by Cambridge University Press

The joint LR test of significance of the four age adjustment parameters gives an approximate χ^2_4 variate of $-2(LL_6 - LL_1) = 8.98, 0.05 < P < 0.10$. The age dependence is no doubt masked by the nonuniform age distribution of our sample; 50% are aged between 20 and 30 years while relatively few are older than 50 or younger than 20 years.

The curves in Figures 1 and 2, derived from the maximum likelihood estimates in Table 3, are given by $y = \log_e \{\hat{\Pi}_o/(1-\hat{\Pi}_o)\} + \hat{\alpha}t + \hat{\beta}t^2$. They appear to fit the data well when compared with the observed proportions of drinkers across age groups.

Comparison With Liability Threshold Method for Analysis of Twin Data

The raw data (Table 2) have also been analysed using the liability threshold method [11]; the estimated proband concordance rates (P_c), the correlation coefficient of liability (r) and its standard error ($V_r^{1/2}$) are given in Table 5. The correlation coefficients calculated by this method are all greater than those of Table 4, and although the two sets of correlation coefficients provide consistent summaries of the data, they have different interpretations (see Discussion).

DISCUSSION

Our method for the analysis of twin concordance rates for a binary trait has several obvious advantages. Firstly, it can allow for the effects of confounding factors, such as age and sex, that might otherwise cause the twin correlation estimates to be inflated for reasons which cannot be attributed to either genetic or environmental causes. Secondly, as the model is formulated in terms of the sample likelihood, likelihood ratio criteria can be used to assess the need for additional parameters in the model.

Kaprio et al [6] have also considered the analysis of twin concordance data, but only for explanatory or confounding variables which are qualitative. Their work is useful, but suffers because they use a definition of concordance which ignores the information provided by discordant twin pairs, secondly because they have not extended their model to allow for the confounding effects of quantitative variables, and thirdly because they have not developed their models in terms of parameters which have a direct biological interpretation.

It is of particular importance to compare our likelihood method with the liability threshold model as described by Smith [11]. That model is founded upon the dubious assumption of bivariate normality in liability. Smith [11] defends this approach by arguing, correctly, that any continuous liability distribution can be transformed to normality by an appropriate transformation of scale. However, marginal normality does not always imply bivariate normality [7], which is a much stronger assumption. Moreover, in this context the assumption is untestable. Thus, using the liability model, it is not obvious that the correlation coefficient in liability between relatives can be interpreted, as Smith suggests, in terms of the additive effects of genetic and environmental components of variance.

In contrast, our more direct method avoids the strong and questionable assumption that there is an underlying liability with a bivariate normal distribution. Our parameterisation is in terms of the marginal and joint probabilities that one or both twins are affected. Accordingly, our derived correlation coefficient, ρ, cannot be directly equated with the "correlation coefficient in liability between relatives," r [11]. Although ρ cannot be interpreted as a simple sum of genetic and environmental components as has been suggested for r in the liability model, nevertheless, it is possible to compare the correlation ρ for MZ and DZ twins using this new model and thus to obtain at least qualitative

TABLE 6. Estimates of Coefficient of Genetic Determination ($\hat{G}$)

		$\hat{G}$	SE
Liability Threshold model	Males	1.34	1.12
	Females	0.41	1.14
Likelihood Model	Males	1.02	0.44
	Females	0.18	0.41

information regarding the importance of genetic factors. Further extensions are planned to allow the possible effects of environmental and genetic factors on ρ to be assessed in quantitative terms (See Appendix A).

A comparison of Table 4 with Table 5 shows that, for the present data set, the estimates of r (liability model) are consistently greater than for ρ (likelihood model). More importantly, the standard errors of the r estimates are greater, both in an absolute and proportional sense, than those of the ρ estimates. This simple observation provides a statistical justification for the use of the likelihood method on the grounds of improved efficiency.

On the other hand, as ρ and r relate to different entities, it can be argued that they should not be compared on statistical grounds alone, but also on the basis of utility. For example, as r can (arguably) be decomposed into additive genetic and environmental components, it can also be used to calculate confidence limits for the coefficient of genetic determination (Table 6) as suggested by Smith [11]. As the scale for ρ is different, and nonadditive, it would be impossible to interpret, in any rigorous quantitative way, a "coefficient of genetic determination" defined as twice the difference between ρ_{mz} and ρ_{dz}. This might provide one utilitarian justification for the continued use of r rather than ρ. The strength of this conclusion depends on the validity of the assumptions underlying the liability method, and the credence attached to the global concepts of "heritability" and "genetic determination". It is our hope that by extending the methods introduced in this paper, we will be able to test some of the implicit assumptions about the ways in which genetic and environmental factors interact to influence binary traits in twins. Such extensions could allow inferences to be made which are stronger than those which depend on uncritical application of the liability model, where some of the underlying assumptions cannot be tested.

The limited data presented in this paper are used to illustrate the methods of statistical analysis, rather than to arrive at conclusions about the causes of alcohol use in twins. Nevertheless, as indicated above, the data suggest that (a) there is bias in the ascertainment of male DZ twins, and (b) the concordance rates for alcohol use for MZ twin pairs are somewhat greater than for DZ twin pairs. These conclusions will be tested with data on alcohol consumption from a much larger sample of adult twin pairs.

ACKNOWLEDGMENTS

This work was supported by the Australian Associated Brewers, the Australian Tobacco Research Foundation, the National Health and Medical Research Council, and the Victor Hurley Fund of the Royal Melbourne Hospital. We thank Jan Temperley, Claire Thomson and Elizabeth Walpole for their valuable help with this study.

REFERENCES

1. Baker RJ, Nelder JA (1978): "The GLIM System. Release 3." Oxford: Numerical Algorithms Group.
2. Box GEP, Cox DR (1964): An analysis of transformations. J Royal Stat Soc, Series B. 26:211–252.

3. Cox DR (1970): "Analysis of Binary Data." London: Methuen, pp 14–29, 61–72, 87–100.
4. Cox DR, Hinkley DV (1974): "Theoretical Statistics." London: Chapman and Hall, pp 279–363.
5. Kaplan B, Elston RC (1972): A subroutine package for maximum likelihood estimation (MAXLIK). Chapel Hill: University of North Carolina, Institute of Statistics Mimeo Series No. 823.
6. Kaprio J, Sarna S, Koskenvico M (1981): Multivariate logit analysis of concordance ratios for qualitative traits in twin studies. Acta Genet Med Gemellol 30:267–274.
7. Karlin S (1979): Comments on statistical methodology in medical genetics. In Sing CF, Skolnick M (eds): "Genetic Analysis of Common Diseases: Applications to Predictive Factors in Coronary Disease." New York: Alan R. Liss, pp 497–520.
8. Kendall M, Stuart A (1977): "The Advanced Theory of Statistics." Vol. 1. London: Griffin, pp 243–262.
9. Hopper JL, Mathews JD (1982): Extensions to multivariate normal models for pedigree analysis. Ann Hum Genet 46:373–383.
10. Landis JR, Koch GG (1977): A one-way components of variance model for categorical data. Biometrics 33:671–679.
11. Smith C (1974): Concordance in twins: Methods and interpretation. Am J Hum Genet 26:454–466.

Correspondence: Dr. J.D. Mathews, University of Melbourne, Department of Medicine, Royal Melbourne Hospital, Victoria 3050, Australia.

APPENDIX A

The within pair correlation ρ, like Π, may depend upon explanatory variables. In particular, where the correlation depends upon shared environmental factors we may wish to model ρ as a function of duration of cohabitation or of time since separation at time t_o [9]. For example, we may write (see Fig. 3)

$$\rho = \rho_0 + \rho_1 e^{-\nu \max(t-t_0,0)}$$

where t is age, ρ_0 might represent genetic and/or constant environmental correlation, ρ_1 the correlation due to changing shared environment, and ν the attentuation rate of this correlation with time.

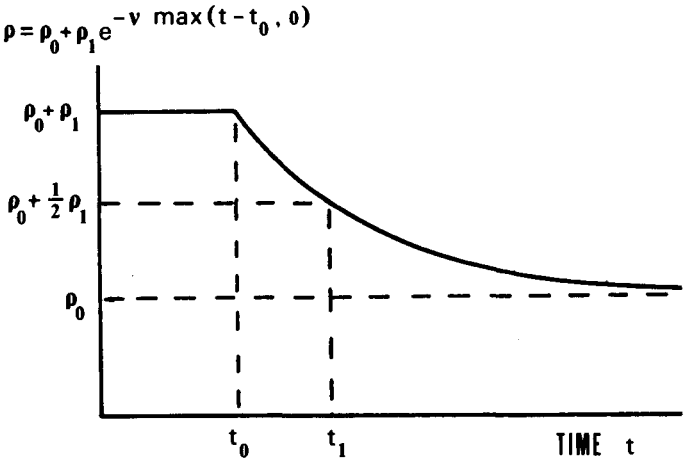

Fig. 3. Correlation, ρ, as a function of time (or age), t, with exponential decay after twin separation at time t_0, where ρ_0 and ρ_1 are interpreted as components of correlation and ν controls the rate of attenuation of ρ_1 after t_0. $t_1 = t_0 + \nu^{-1}\log_e 2$.

https://doi.org/10.1017/S0001566000006413 Published online by Cambridge University Press

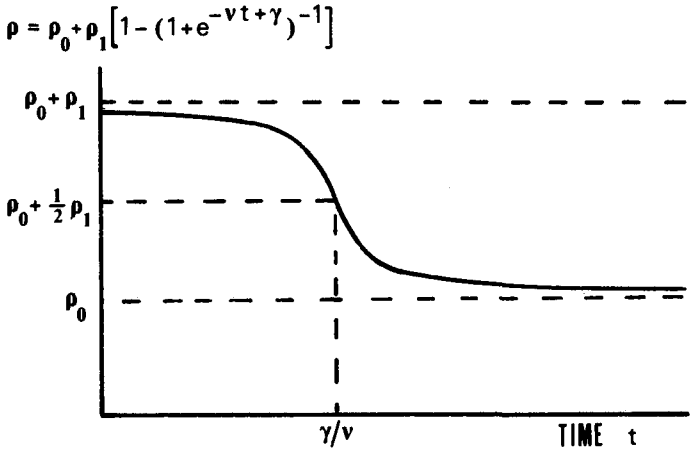

Fig. 4. Correlation, ρ, as a logistic function of time: ρ_0 and ρ_1 are interpreted as components of correlation, and ν and γ control the rate and timing of attenuation of ρ_1.

Alternatively, write $\rho = \rho_0 + \rho_1\{1 - (1 + e^{-\nu t+\gamma})^{-1}\}$ (see Fig. 4), where ρ_0 and ρ_1 have interpretations as above, and ν and γ control the rate and timing of attenuation in the correlation due to shared environment. Under the assumption of homogeneity of environment ρ_1 should be the same for MZ and DZ twins, and in theory this could be used to test the assumption. Depending upon the influence of genes, ρ_0 for MZ twins (ρ_{omz}) should be greater than or equal to its value for DZ twins (ρ_{odz}). However, unlike the genetic components of correlation in the liability model, ρ_{omz} is not necessarily equal to twice ρ_{odz} even if the trait is exclusively genetic in origin. Many other different formulations for ρ are possible.

APPENDIX B

Plots of the proportion of drinkers against age (Figs. 1 and 2) reveal an asymmetry which would lead to systematic bias if a simple quadratic function were used to adjust for age dependence. It is possible that the age adjustment could be improved by an initial transformation (eg, log) of the age scale. For greater flexibility and utility we suggest a generalized power transformation, chosen according to some goodness of fit criterion. Following Box and Cox [2], define the transformed (age) variable

$$t^{(\lambda)} = \begin{cases} \dfrac{t^{\lambda}-1}{t} & \text{if } \lambda \neq 0, \\ \log_e t & \text{if } \lambda = 0. \end{cases}$$

This family of power transformations is continuous in λ at 0, and therefore can be incorporated into the log likelihood function without causing a singularity in the likelihood surface.

For our alcohol data, t was taken to be (age in years $-$ 10)/10 and maximization of the likelihood of the saturated model (Model 1) with an age transformation parameter for each sex yields the following estimates:

$$\lambda_m = 0.46 \text{ (SE} = 1.65), \lambda_f = 0.32 \text{ (SE} = 0.82)$$

https://doi.org/10.1017/S0001566000006413 Published online by Cambridge University Press

with LL = −183.99. Setting $\lambda_m = \lambda_f = 1$ yields LL = −184.20. Clearly the estimates for λ_m and λ_f are not significantly different from 0 nor 1, and the change in the log likelihood is negligible indicating only slight improvement of fit for the inclusion of two extra parameters.

For the purposes of the analysis of the alcohol data, the transformation parameter was set, somewhat arbitrarily, to 0.33, corresponding to a cube root transformation of the age scale, for both sexes.

APPENDIX C

For an alternative but mathematically equivalent formulation, the model may be expressed in terms of Π and the proband concordance rate $P_c = \Pr\{X_2 = 1 \mid X_1 = 1\}$. From (2)

$$\rho = (P_c - \Pi) / (1 - \Pi). \tag{6}$$

Let us consider just one sex by zygosity twin class. Under the simplest model in which Π, ρ and P_c are independent of all explanatory variables suppose we observe x_{11} concordant positive pairs, x_{00} concordant negative pairs, and x_d discordant pairs. The log likelihood in terms of P_c and Π is:

$$LL = x_{11}\log_e P_c + x_{11}\log_e \Pi + x_d \log_e (1-P_c) + x_d \log_e \Pi + x_{00}\log_e \{1-\Pi(2-P_c)\},$$

and when maximized analytically gives the MLEs

$$\hat{\Pi} = (2x_{11} + x_d)/2N \text{ and } \hat{P}_c = 2x_{11}/(2x_{11} + x_d) \text{ (c.f. [11])}.$$

The observed inverse information matrix at $\hat{\theta} = (\hat{\Pi}, \hat{P}_c)$ is

$$I^{-1} = \left(\frac{-\partial^2 LL}{\partial\theta_i \partial\theta_j}\right)\Bigg|_{\hat{\theta}}^{-1} = \begin{pmatrix} a & b \\ b & c \end{pmatrix}^{-1} \quad \text{where}$$

$$a = 4N^3 (x_{11} + x_d)/\{x_{00} (2x_{11} + x_d)^2\},$$
$$b = -N^2/x_{00},$$
$$c = (2x_{11} + x_d)^2 \{(4x_{11})^{-1} + x_d^{-1} + (4x_{00})^{-1}\}.$$

Let $d = \det(I)^{-1} = (ac - b^2)^{-1}$, then

$$\text{Var}\,\hat{\Pi} \simeq cd, \text{ Var } \hat{P}_c \simeq ad \text{ and } \text{Cov}(\hat{\Pi}, \hat{P}_c) \simeq -bd$$

are asymptotic estimates of variance and covariance [4].

From (6) the MLE for ρ is $\hat{\rho} = (\hat{P}_c - \hat{\Pi})/(1 - \hat{\Pi})$. It can be shown that $\hat{\rho} = (X^2/N)^{1/2}$ where X^2 is Pearson's chi-squared statistic for the two by two table $\{x_{ij}: i,j = 0,1\}$ where $x_{01} = x_{10} = 1/2x_d$. Thus $N\hat{\rho}^2$ provides a simple approximate χ^2_1 test of the hypothesis H_o: $\rho = 0$. Also from (6) we have

$$\text{Var}\,\hat{\rho} \simeq (1-\hat{\Pi})^{-2}\{K^2 \text{Var}\,\hat{\Pi} + \text{Var}\,\hat{P}_c - 2K\,\text{Cov}(\hat{\Pi}, \hat{P}_c)\},$$

where $K = (1-\hat{P}_c)/(1-\hat{\Pi})$, [5,8].

https://doi.org/10.1017/S0001566000006413 Published online by Cambridge University Press