Biometrika (1992), 79, 3, *pp.* 441-61
Printed in Great Britain

Response models for mixed binary and quantitative variables

BY D. R. COX
Nuffield College, Oxford OX1 1NF, U.K.

AND NANNY WERMUTH
Psychological Institute, University of Mainz, 6500 Mainz, Germany

SUMMARY

A number of special representations are considered for the joint distribution of qualitative, mostly binary, and quantitative variables. In addition to the conditional Gaussian models and to conditional Gaussian regression chain models some emphasis is placed on models derived from an underlying multivariate normal distribution and on models in which discrete probabilities are specified linearly in terms of unknown parameters. The possibilities for choosing between the models empirically are examined, as well as the testing of independence and conditional independence and the estimation of parameters. Often the testing of independence is exactly or nearly the same for a number of different models.

Some key words: Conditional Gaussian model; Graphical chain model; Linear model; Logistic function; Multivariate normal distribution; Probit model.

1. INTRODUCTION

The object of this paper is to compare a number of models for the joint distribution of quantitative and binary response variables. One role of such models is as a route for testing hypotheses of independence or conditional independence. We examine the extent to which essentially the same test arises from different models. A further important point is that for some models particular null hypotheses may be satisfied only under much stronger versions of independence than those it is desired to test, so that the models are unsuitable for the required purpose.

Two of the families of models under consideration are models based on conditional Gaussian distributions, i.e. for conditional normality of the continuous components and an arbitrary distribution for the discrete components, and on conditional Gaussian regressions. These take some conditional regression relations from a conditional Gaussian distribution and then separately assign distributions, possibly arbitrary, to the conditioning variables. The latter have been introduced as graphical chain models by Lauritzen & Wermuth (1989). In addition we examine some aspects of models in which the probabilities for the binary components are specified linearly and also of models in which there is an initial multivariate normal distribution from which the binary components, or more generally the ordinal components, are derived by forming discrete classes. Indeed the connection between multivariate continuous distributions, in particular the multivariate normal distribution, and binary and ordinal data has a long history and many facets. Probit-style models (Finney, 1952) for binary variables generated from a normal distribution of underlying observed 'tolerances' form probably the most familiar example. Another instance is Pearson's (1901) tetrachoric correlation in which the relation between two binary variables is summarized via a bivariate normal distribution fitted to the contingency table formed from the discrete responses.

We shall study first in two, then in three, dimensions aspects of the following: the general nature of the relationships between various models, the implications for estimation, and the implications for testing null hypotheses of independence or conditional independence.

2. Bivariate distributions

2·1. *Some distributional results*

We consider first just two response variables, i.e. we focus on the joint distribution of two random variables. It is convenient to separate the discussion into the study of a continuous response conditional on a discrete explanatory variable and an analysis the other way round. This is not a conventional study of dependence, where, even if the explanatory variable is random, conditioning on the observed values of the explanatory variable is used in inference.

Suppose first that (U, X) are bivariate normal with zero means, unit variances and correlation ρ_{xu}. Let a dichotomous variable A be formed from U via a cut-off point α; we write

$$A=\begin{cases}1 & (U \geqslant \alpha),\\ 0 & (U<\alpha).\end{cases} \qquad (2\cdot1)$$

Then derivatives of the moment generating function of a truncated normal distribution (Tallis, 1961) or direct calculations show that

$$\begin{aligned}\mu_u^+(\alpha)&=E(U\,|\,A=1)=\phi(\alpha)/\Phi(-\alpha),\\ \mu_u^-(\alpha)&=E(U\,|\,A=0)=-\phi(\alpha)/\Phi(\alpha),\end{aligned} \qquad (2\cdot2)$$

where $\phi(.)$ and $\Phi(.)$ are respectively the standard normal density and integral. Further

$$\begin{aligned}\sigma_{uu}^+(\alpha)&=\mathrm{var}\,(U\,|\,A=1)=1+\alpha\mu_u^+(\alpha)-\{\mu_u^+(\alpha)\}^2,\\ \sigma_{uu}^-(\alpha)&=\mathrm{var}\,(U\,|\,A=0)=1+\alpha\mu_u^-(\alpha)-\{\mu_u^-(\alpha)\}^2,\end{aligned} \qquad (2\cdot3)$$

and the third cumulants, or third moments about the mean, are

$$\kappa_{3,u}^+(\alpha)=(\alpha^2-1)\mu_u^+(\alpha)-3\alpha\mu_u^+(\alpha)^2+2\mu_u^+(\alpha)^3,$$

with a corresponding formula for $\kappa_{3,u}^-(\alpha)$, and from these the standardized third cumulants

$$\gamma_{1,u}^+(\alpha)=\kappa_{3,u}^+(\alpha)/\{\sigma_{uu}^+(\alpha)\}^{3/2} \qquad (2\cdot4)$$

and $\gamma_{1,u}^-(\alpha)$ are calculated directly.

Because $X=\rho_{xu}U+\varepsilon_{x.u}$, where $E(\varepsilon_{x.u})=0$, $\mathrm{cov}\,(U, \varepsilon_{x.u})=0$, $\mathrm{var}\,(\varepsilon_{x.u})=\sigma_{xx.u}=1-\rho_{xu}^2$, we have that

$$\mu_{x.a}^+=E(X\,|\,A=1)=\rho_{xu}\mu_u^+(\alpha),\quad \sigma_{xx.a}^+=\mathrm{var}\,(X\,|\,A=1)=\rho_{xu}^2\sigma_{uu}^+(\alpha)+(1-\rho_{xu}^2),$$

with corresponding formulae for $\mu_{x.a}^-$, $\sigma_{xx.a}^-$ obtained by replacing + by − everywhere. Further

$$\kappa_{3,x.a}^+=E\{(X-\mu_{x.a}^+)^3\,|\,A=1\}=\rho_{xu}^3\kappa_{3,\mu}^+(\alpha).$$

Note that if U is not observed there is no loss of generality in taking it in standardized form. If Y has mean μ_y and variance σ_{yy}, the above formulae for X correspond to the moments of $(Y-\mu_y)/\surd\sigma_{yy}$ so that in this general case

$$\mu^+_{y.a}=\mu_y+\rho_{yu}\sigma^{\frac{1}{2}}_{yy}\mu^+_u(\alpha),\quad \sigma^+_{yy.a}=\sigma_{yy}\{\rho^2_{yu}\sigma^+_{uu}(\alpha)+(1-\rho^2_{yu})\},$$
$$\kappa^+_{3,y.a}=\sigma^{3/2}_{yy}\rho^3_{yu}\kappa^+_{3,u}(\alpha). \tag{2·5}$$

For studying dependencies of A on Y we start from the conditional distribution of U given $Y=y$ which is normal with mean $\rho_{yu}(y-\mu_y)/\surd\sigma_{yy}$ and variance $(1-\rho^2_{yu})$ so that

$$\text{pr}\,(A=1\mid Y=y)=\Phi\left\{\frac{\rho_{yu}(y-\mu_y)/\surd\sigma_{yy}-\alpha}{\surd(1-\rho^2_{yu})}\right\}. \tag{2·6}$$

In many practical cases this will be virtually indistinguishable from a linear logistic regression (Cox, 1966; Cox & Snell, 1989, p. 22), i.e. the simplest form of a conditional Gaussian regression with discrete response, and provided the probabilities are not too extreme, say $0{\cdot}2\leqslant \text{pr}\,(A=i\mid Y=y)\leqslant 0{\cdot}8$, this implies near linearity of the log odds, i.e.

$$\log\frac{\text{pr}\,(A=1\mid Y=y)}{\text{pr}\,(A=0\mid Y=y)}\simeq d\frac{\rho_{yu}(y-\mu_y)/\surd\sigma_{yy}-\alpha}{\surd(1-\rho^2_{yu})}, \tag{2·7}$$

where the most suitable value of the constant d depends on the range over which the approximation is required. To match the functions at the 20%, 50% and 80% points, we take $d=1{\cdot}65$.

By contrast, in the conditional Gaussian distribution the conditional distributions of Y given $A=0, 1$ are normal. In the homogeneous case, the conditional variances are the same at both levels of A, while the marginal probabilities $\text{pr}\,(A=i)$ are positive but otherwise arbitrary. The relationship $\text{pr}\,(A=1\mid Y=y)$ derived from this joint distribution is linear logistic in the homogeneous and quadratic logistic in the nonhomogeneous case, while the marginal distribution of Y is a mixture of normals. Any conditional Gaussian regression looks like a conditional distribution derived from a joint conditional Gaussian distribution: with a dichotomous response, A, it is a logistic regression and with a continuous response, Y, it is a linear regression. A joint distribution defined by a sequence of conditional Gaussian regressions and a marginal conditional Gaussian distribution of the variables which are not responses is called a conditional Gaussian regression chain model. With just two variables, A and Y, the conditional Gaussian regression chain model with Y as response to A defines a joint conditional Gaussian distribution while, in general, this is not the case for a conditional Gaussian regression chain model with A as response to Y.

Both the conditional Gaussian regression chain model for A as response to Y and the dichotomized normal model are special cases of one in which Y is marginally normal and in which

$$\pi^{A|Y}_{1|y}=\text{pr}\,(A=1\mid Y=y)=G(\theta^{(g)}_0+\theta^{(g)}_1 y), \tag{2·8}$$

where $G(x)$ is the logistic function $L(x)=e^x/(1+e^x)$ for the conditional Gaussian regression chain model, and $G(x)$ is the standardized normal distribution function $\Phi(x)$ for the dichotomized normal. Clearly other choices of G are possible, conceivably containing additional nuisance parameters to allow a data-based choice of G (Aranda-Ordaz, 1981). Despite some obvious limitations, the choice $G(x)=x$, leading to a linear model for probabilities, is useful in particular as an approximation to the logistic and probit functions wherever, as noted above, the conditional probabilities are largely confined to the range (0·2, 0·8). We therefore write

$$\pi^{A|Y}_{1|y} = \pi^A_1 + \beta_{ay}(y-\mu_y), \quad \pi^{A|Y}_{0|y} = \pi^A_0 - \beta_{ay}(y-\mu_y), \tag{2·9}$$

provided that the values so defined are in [0, 1]. In some applications the assumption that the probabilities are confined to a central range is entirely reasonable. An example are probabilities of changing a field of study investigated by Weck (1991) under a wide range of different conditions in Germany; see § 3·2.

One advantage of the linear representation is the very direct interpretation of the parameters and another is that marginalization over Y is immediate. Provided only that $E(Y)=\mu_y$, we recover the stated marginal probability. Under the dichotomized normal model, the corresponding probability is $\Phi(-\alpha)$, whereas under the conditional Gaussian regression chain model the marginal probability is $E_Y\{L(\theta_0^{(\lambda)}+\theta_1^{(\lambda)}Y)\}$, where the expectation is over the normal distribution of Y, $N(\mu_y, \sigma_y^2)$. While this cannot be evaluated exactly in closed form, a good approximation is obtained by writing $L(x) \simeq \Phi(xc)$, where $c=0{\cdot}607$. Then, on omitting the superscript λ for convenience, we have

$$E\{L(\theta_0+\theta_1 y)\} \simeq \int_{-\infty}^{\infty} \Phi(c\theta_0+c\theta_1 y)\phi(y;\mu_y,\sigma_y^2)\,dy,$$

where $\phi(y;\mu_y,\sigma_y^2)$ is the density of $N(\mu_y,\sigma_y^2)$. This integral can be evaluated in closed form to give

$$\Phi\left\{\frac{c(\theta_0+\theta_1\mu_y)}{\surd(1+c^2\theta_1^2\sigma_y^2)}\right\} \simeq L\left\{\frac{\theta_0+\theta_1\mu_y}{\surd(1+c^2\theta_1^2\sigma_y^2)}\right\}. \tag{2·10}$$

The advantages of the simpler marginalization of (2·9) are particularly strong in a larger number of dimensions.

A model in which, for example, the conditional distribution of Y given $A=i$ is normal and the marginal distribution of Y therefore nonnormal is distinct from a model in which the marginal distribution is normal. Nevertheless the separation of the two normal components in the first model must be appreciable if the distinction is to be detectable with realistic amounts of data. This can be verified numerically or seen analytically by noting that a mixture with probabilities $(\frac{1}{2}+\frac{1}{2}\xi)$, $(\frac{1}{2}-\frac{1}{2}\xi)$ of normal distributions of means δ, $-\delta$ and unit variances has density

$$\begin{aligned} f_y^Y &= \frac{1}{\surd(2\pi)}\{\tfrac{1}{2}(1+\xi)\,e^{-\frac{1}{2}(y-\delta)^2}+\tfrac{1}{2}(1-\xi)\,e^{-\frac{1}{2}(y+\delta)^2}\} \\ &= \frac{1}{\surd\{2\pi(1+\delta^2-\xi^2\delta^2)\}}\exp\{-\tfrac{1}{2}(1+\delta^2-\xi^2\delta^2)^{-1}(y-\xi\delta)^2\}\{1+O(\delta^3)\}, \end{aligned} \tag{2·11}$$

after some manipulation, thus showing that nonnormality enters only in the term of order δ^3. A similar argument shows that if the marginal distribution of Y is normal and the relation between A and Y probit or logistic confined to the range (0·2, 0·8) and hence effectively linear, then the conditional distribution of Y given $A=i$ is very close to normality. To see this analytically note that, with $i^*=2i-1$,

$$f^{Y|A}_{y|i} = f^Y_y f^{A|Y}_{i|y}/f^A_i \simeq \frac{1}{\surd(2\pi)\sigma_y}\exp\left\{-\frac{(y-\mu_y)^2}{2\sigma_y^2}\right\}\{f_i^A+\gamma_{ay}(y-\mu_y)i^*\}/f_i^A. \tag{2·12}$$

The limitation on the probabilities implies that $\gamma_{ay}\sigma_y$ is small. If we write $\gamma_{ay} = \varepsilon/\sigma_y$ and incorporate the linear term into the exponential we have that with an error of order ε^3 the conditional distributions are normal with the same variance.

2·2. *General statistical interpretation*

A number of broad conclusions can be drawn from the above results, in particular from the numerical results in Tables 1 and 2 which are based directly on the formulae of § 2·1. By the symmetry of the problem it is enough to suppose that $\alpha \geqslant 0$.

Table 1. *Variance ratios $\sigma^+_{yy.a}/\sigma^-_{yy.a}$ of Y given $(A = i)$ in a dichotomized bivariate normal distribution*; α, *cut-off point for dichotomized U*; ρ_{yu}, *correlation*

α	$\rho_{yu} = 0{\cdot}2$	$\rho_{yu} = 0{\cdot}5$	$\rho_{yu} = 0{\cdot}8$	$\rho_{yu} \simeq 0{\cdot}99$
0	1·000	1·000	1·000	1·000
0·5	0·991	0·938	0·639	0·552
1·5	0·975	0·835	0·533	0·193
2·5	0·965	0·781	0·429	0·093

Table 2. *Standardized skewnesses $\gamma^+_{1y.a}$ and $\gamma^-_{1y.a}$ of Y given $(A = i)$ in a dichotomized bivariate normal distribution*; α, *cut-off point*; ρ_{yu}, *correlation*

α	Level of A	$\rho_{yu} = 0{\cdot}2$	$\rho_{yu} = 0{\cdot}5$	$\rho_{yu} = 0{\cdot}8$	$\rho_{yu} \simeq 0{\cdot}99$
0·0	+	0·002	0·035	0·245	0·995
	−	−0·002	−0·035	−0·245	−0·995
0·5	+	0·001	0·028	0·215	1·169
	−	−0·002	−0·042	−0·252	−0·800
1·5	+	0·001	0·015	0·139	1·444
	−	−0·002	−0·036	−0·172	−0·391
2·5	+	0·000	0·008	0·084	1·676
	−	−0·001	−0·012	−0·051	−0·102

If emphasis is on the conditional distribution of the continuous component Y given the discrete component A, then, under the model based on a bivariate normal distribution, unequal variances combined with skewness will be encountered in the two groups if α and ρ_{yu} are not equal to zero. These effects are likely to be important and empirically detectable only if the correlation between Y and U is fairly high, and, for the inequality of variances, if the dichotomy of U is into quite unequal groups. By contrast in the conditional Gaussian distribution model, the conditional distributions of Y given A are normal and in its homogeneous form have equal variances.

In the dichotomized normal model the conditional distribution of A given Y has probit form: clear departure from that would be evidence against an underlying bivariate normal distribution. The joint distribution of (A, Y) has four independent parameters which can be taken in various forms; $(\mu_y, \sigma_{yy}, \rho_{yu}, \alpha)$ is one natural choice. The parameters can be estimated in several ways; see § 2·3.

Broadly similar results apply when U is divided into three groups. If the trichotomy is symmetrical, the variance of Y within groups will be the same in the outer groups and different, in general, for the central group. Furthermore, the conditional distributions of

A given Y will have probit form in the outer groups but not for the central group. As the number of groups increases we quite rapidly approach recovery of the information about correlation that would be available were the underlying continuous variables to the observed (Cox, 1958).

2·3. *Estimation*

The estimation of parameters from a conditional Gaussian distribution model and from a conditional Gaussian regression chain model with discrete response follows standard maximum likelihood methods. For n independent observations $(i_1, y_1), \ldots, (i_n, y_n)$ from the dichotomized bivariate distribution, where $i_r = 1$ if $U_r > \alpha$ and $i_r = 0$ if $U_r \leqslant \alpha$, the log likelihood is best written as the sum of the marginal log likelihood from $(y_1, \ldots, y_n)$ and the log conditional likelihood for A given $Y = y$ writing $\operatorname{pr}(A = 1 \mid Y = y)$ in the form $\Phi(\theta_0 + \theta_1 y)$, where

$$\theta_0 = -(\rho_{yu}\mu_y/\sigma_{yy}^{\frac{1}{2}} + \alpha)(1-\rho_{yu}^2)^{-\frac{1}{2}}, \quad \theta_1 = \rho_{yu}\{\sigma_{yy}(1-\rho_{yu}^2)\}^{-\frac{1}{2}}.$$

Thus, $\hat{\mu} = \bar{y}$, $\hat{\sigma}_{yy} = \Sigma\,(y_r - \bar{y})^2/n$ from marginal normality and they are asymptotically uncorrelated with $\hat{\theta}_0$, $\hat{\theta}_1$ derived via a probit analysis (Maritz, 1953) of $(i_1, \ldots, i_n)$ on $(y_1, \ldots, y_n)$.

An older and computationally simpler method proceeds by first noting that because $\operatorname{pr}(U > \alpha) = \operatorname{pr}(A = 1) = \Phi(-\alpha)$ we can estimate α by $\tilde{\alpha} = \Phi^{-1}(\bar{i})$, where $\bar{i}$ is the overall proportion of 1's. Further

$$E(Y \mid A = 1) - E(Y \mid A = 0) = \rho_{yu}\sigma_{yy}^{\frac{1}{2}}\phi(\alpha)\{\Phi(\alpha)\Phi(-\alpha)\}^{-1}, \qquad (2·13)$$

so that in a self-explanatory notation we can write

$$\tilde{\rho}_{yu} = \frac{(\bar{Y}^{+} - \bar{Y}^{-})\bar{i}(1-\bar{i})}{\hat{\sigma}_{yy}^{\frac{1}{2}}\phi\{\Phi^{-1}(\bar{i})\}}, \qquad (2·14)$$

where $\hat{\sigma}_{yy} = \Sigma\,(y_r - \bar{y})^2/n$. This is Pearson's (1903) biserial correlation coefficient. The dependence of $\tilde{\rho}_{yu}$ on $\bar{i}$ near $\bar{i} = \frac{1}{2}$, i.e. median dichotomy, is very slow; as $\bar{i}$ varies from 0·5 to 0·3 the $\bar{i}$-dependent factor in (2·14) changes only from $0·624 = \surd(\pi/8)$ to 0·604. Then, for some purposes it is sensible to replace (2·14) by

$$\tilde{\rho}_{yu}^{*} = \frac{(\bar{Y}^{+} - \bar{Y}^{-})}{\hat{\sigma}_{yy}^{\frac{1}{2}}}\sqrt{\left(\frac{\pi}{8}\right)}. \qquad (2·15)$$

Comparison of the asymptotic variances of $\tilde{\rho}_{yu}$ and $\hat{\rho}_{yu}$ shows that the asymptotic efficiency of $\tilde{\rho}_{yu}$ relative to $\hat{\rho}_{yu}$ is 1·00 if $\rho = 0$, a result related to a general result about testing independence to be discussed in § 2·4. The efficiency is, however, appreciably less than one when $\rho > \frac{1}{2}$, say (Tate, 1955). For this result one needs the asymptotic variance of $\tilde{\rho}_{yu}$ (Soper, 1915), calculated most directly by finding the asymptotic covariance matrix of $\Sigma\, I_r$, $\Sigma\, Y_r$, $\Sigma\, I_r Y_r$, $\Sigma\,(Y_r - \bar{Y})^2$ in terms of which $\tilde{\rho}_{yu}$ can be expressed. Note that to the required accuracy $\Sigma\,(Y_r - \bar{Y})^2$ can be replaced by $\Sigma\,(Y_r - \mu_y)^2$. The asymptotic variance of $\hat{\rho}_{yu}$ is obtained from the Fisher information matrix (Tate, 1955; Prince & Tate, 1966).

Fitting of the model linear in probabilities (2·9) is most conveniently done by unweighted least squares applied to the (0, 1) responses, comparing the residual mean square to $\bar{i}(1-\bar{i})$ for an approximate test of adequacy. It would also be possible to fit this model by maximum likelihood. Approximate calculations in Appendix 1 imply that, provided the fitted probabilities lie in the range (0·2, 0·8), the point estimates will be nearly the same and the reduction in variance is small, at most 5% and usually much less than this.

An example in which least squares and maximum likelihood estimation show these properties is given with the following data, collected by Dr N. Schmitt in connection with a medical dissertation at the pain clinic in Mainz. Success of treatment is predicted from a score for stage of chronic pain for $n = 58$ male patients treated for three weeks in the pain clinic. Table 3 shows predicted probabilities of successful treatment under a logit, a probit and a linear-in-probabilities regression. The only notable difference is that, in fitting a linear model by maximum likelihood, relatively greater weight is attached to the two individuals at the extreme level of chronic pain $y = 11$.

Table 3. *Different estimates for regression of treatment success on stage of chronic pain, y*

			Probabilities of treatment success, estimated by				
		Number	observed relative	linear	linear	linear regression in	
	Total	of	frequen-	logistic	probit	binary responses via	
Stage	count	successes	cies	regression	regression	LS	ML
y	n_y	n_{1y}	n_{1y}/n_y	(a)	(b)	(c)	(d)
6	8	7	0·88	0·75	0·76	0·75	0·78
7	9	5	0·56	0·64	0·64	0·63	0·64
8	15	6	0·40	0·50	0·50	0·50	0·51
9	14	6	0·43	0·36	0·36	0·37	0·37
10	10	3	0·30	0·24	0·24	0·24	0·23
11	2	0	0·00	0·15	0·15	0·11	0·01

(a) $\log(\hat{\pi}_{1|y}/\hat{\pi}_{0|y}) = 4{\cdot}52 - 0{\cdot}57y$, (b) $\Phi^{-1}(\hat{\pi}_{1|y}/\hat{\pi}_{0|y}) = 2{\cdot}82 - 0{\cdot}35y$,

(c) $\hat{\beta}_{\mathrm{LS}} = (1{\cdot}52, -0{\cdot}13)$, $\operatorname{cov}(\hat{\beta}_{\mathrm{LS}}) = \begin{pmatrix} 0{\cdot}1205 & -0{\cdot}0140 \\ \cdot & 0{\cdot}0017 \end{pmatrix}$

(d) $\hat{\beta}_{\mathrm{ML}} = (1{\cdot}61, -0{\cdot}14)$, $\operatorname{cov}(\hat{\beta}_{\mathrm{ML}}) = \begin{pmatrix} 0{\cdot}1170 & -0{\cdot}0136 \\ \cdot & 0{\cdot}0016 \end{pmatrix}$

2·4. *Tests of independence*

In § 2·3 the emphasis is on estimation, for example of the correlation coefficient in an underlying bivariate normal distribution. Sometimes, however, there is special interest in testing the null hypothesis of independence between discrete and continuous components. Here there is a certain robustness to formulation which arises also in purely continuous and purely discrete cases.

For example, the optimal test for independence in a bivariate normal distribution of (X, Y) can be regarded as a test of the correlation coefficient, treating both variables symmetrically, or as a test of linear regression either of Y on X or of X on Y, which in turn can be regarded as arising in a number of ways. Very similar remarks apply to the purely binary case where Fisher's exact test for 2×2 table can be derived from several viewpoints, and analogous results are available for $r \times s$ tables (Birch, 1963).

Faced with a random sample from a mixed binary and continuous distribution, one directly appealing test of independence is based on the difference between the means of the continuous variable Y at the two levels of the binary variable A, standardized similarly to the Student t statistic to produce under the null hypothesis a statistic with approximately a standard normal distribution. Note that the unequal variances examined in § 2·1 arise only when $\rho_{yu}^2 \neq 0$. The resulting test is exactly or asymptotically optimal under the following sets of assumptions.

(i) For each $A = i$, Y is normal with mean μ_i $(i = 0, 1)$ and constant variance, the null hypothesis of independence being equivalent to $\mu_1 = \mu_0$. Here the t test of a difference is directly appropriate.

(ii) To cover several possible types of departure from independence, suppose that the joint distribution of (A, Y) is specified by a marginal density $f_y^Y = g(y;\, \theta)$ for Y and a conditional distribution for A given $Y = y$ of the form

$$\pi_{i|y}^{A|Y} = \mathrm{pr}\,(A = i \mid Y = y) = \{h(\alpha, \beta y)\}^i \{1 - h(\alpha, \beta y)\}^{1-i}, \qquad (2{\cdot}16)$$

where g, h are known functions, and α, β, θ are unknown parameters with independent parameter spaces. The null hypothesis of independence is $\beta = 0$. A crucial point is that y appears linearly in (2·16) multiplying the parameter β, so that probit and logistic regressions are among the many special cases. On evaluating the β-component of the derivative of the log likelihood function at the null hypothesis, we obtain

$$\left[\frac{\partial h(\alpha, \phi)}{\partial \phi}\right]_{\phi = 0} = \left[\sum_{r:i_j = 1} y_r / h(\alpha, 0) - \sum_{r:i_j = 0} y_r / \{1 - h(\alpha, 0)\}\right]. \qquad (2{\cdot}17)$$

Now under the null hypothesis, the maximum likelihood estimate of $h(\alpha, 0)$ is the proportion of observations with $i_r = 1$ so that (2·17) can be replaced by the difference of the two sample means, leading after standardization to the Student t statistic as having the usual asymptotic properties of the score test; see for example Cox & Hinkley (1974, pp. 315, 324). Note that, if y in (2·16) is replaced by suitable nonlinear functions of y, robust tests of location can be generated.

(iii) Finally, note that, under the null hypothesis $\beta = 0$, very generally $(y_1, \ldots, y_n, \Sigma\, i_r)$ or some reduction thereof is sufficient for (α, θ). If then a test is based on $\Sigma\, i_r y_r$, the sample total in the $A = 1$ group, an 'exact' test can be obtained from the permutation distribution of the sample total and will typically be close to the Student t test.

Thus as in the purely continuous or purely binary cases, essentially the same test of independence can be derived from various viewpoints.

2·5. *Bivariate response plus explanatory variable*

Now suppose that, in addition to the bivariate response variable (A, Y), there is on each individual a vector z of explanatory variables which are either not random or, if random, are treated as fixed at their observed values for the purpose of analysis. In the spirit of the previous discussion, some relatively simple models for interpretation of such data are as follows, with $\bar{z}$ denoting the mean of z over the data.

(i) We have that Y given Z is normal with mean $\mu_y + \beta_{yz}^{\mathrm{T}}(z - \bar{z})$, variance $\sigma_{yy.z}$ and, conditionally on $Y = y$ and z, A is governed by a probit law:

$$\mathrm{pr}\,(A = 1 \mid Y = y, z) = \Phi\{\gamma_{a.yz}^{(p)} + \gamma_{ay.z}^{(p)}(y - \mu_y) + \gamma_{az.y}^{\mathrm{T}(p)}(z - \bar{z})\}. \qquad (2{\cdot}18)$$

This is a direct generalization of the model discussed above and can be derived via an underlying bivariate normal distribution of (U, Y) with U dichotomized to form A.

(ii) The probit relation (2·18) can be replaced by a logistic relation. This specifies a homogeneous conditional Gaussian regression chain model if Z has marginally a normal distribution.

(iii) We have that A is governed by a probit law conditionally on z,

$$\text{pr}\,(A=1\,|\,z)=\Phi\{\gamma_{a.z}^{(p)}+\gamma_{az}^{\mathrm{T}(p)}(z-\bar{z})\}, \tag{2·19}$$

and, given $A=i$ and z, Y is conditionally normal with constant variance $\sigma_{yy.za}$ and mean

$$\beta_{y.az}+\beta_{ya.z}i+\beta_{yz.a}^{\mathrm{T}}(z-\bar{z}). \tag{2·20}$$

(iv) The probit relation (2·19) can be replaced by a logistic relation. Normality of Y can then take on at least two forms: (a) if in this case Z is conditionally normal given $A=i$ and has constant variance at both levels of A this specifies a homogeneous conditional Gaussian distribution for A, Y, Z; (b) if, however, Y is marginally normal then a homogeneous conditional Gaussian regression chain model with discrete response results which is different from the one under (ii).

The above models have strong assumptions not only of linearity but, at least as importantly, of parallelism of regression lines. Nonparallelism, at least in extreme cases, can have major substantive implications. Therefore checks of parallelism are necessary in applications.

For instance in (2·20) we may allow $\beta_{yz.a}^{\mathrm{T}}$ to depend on levels of A, for example by inserting $i(z-\bar{z})$ as an additional explanatory variable, possibly constraining the nonparallelism to certain components of z. This gives generalized linear models (McCullagh & Nelder, 1989). If in addition the variance of Y $\sigma_{yy.za}$ is allowed to depend on the level of A, then this specifies not a generalized linear model, but, in the case of marginally normal Y, that is, case (b) under (iv), a nonhomogeneous graphical chain model.

2·6. *Comparison of models*

The results of §§ 2·2 and 2·4 suggest that empirical choice between the various models studied here is likely to be feasible only when substantial correlation is present between the binary and continuous components.The difficulties are likely to be compounded when 'fixed' explanatory variables are present.

A key distinction is between the conditional Gaussian distribution in which for each $A=i$ the continuous variable Y has a 'simple' form and those models in which the marginal distribution of Y is of 'simple' form. Typically, simplicity in the marginal distribution of Y corresponds to fairly complicated conditional distributions and vice versa; see Table 4. However, whenever the linear-in-probability model is a suitable approximation to the conditional dependence of A on Y, then marginal normality of Y corresponds to approximately normal conditional distributions of Y given $A=i$ by (2·12), while conditional normal distributions of Y given $A=i$ correspond to an approximately normal marginal distribution of Y by (2·11). If in this latter case the conditional distribution of A given $Y=y$ is of probit form, it corresponds to an underlying bivariate normal distribution; if instead the conditional distribution of A given $Y=y$ is of logistic form, the specifications correspond to those of a homogeneous conditional Gaussian regression chain model, and, as noted before, these two models are likely to be close.

Dr G. K. Reeves, in work as yet unpublished, has confirmed these qualitative conclusions by imbedding the conditional Gaussian distribution and the conditional Gaussian regression chain models in a single family containing an additional parameter taking values 0 and 1 for the two families in question and showing that the profile likelihood for that parameter is typically very flat.

Table 4. *Distributional properties of models for A, Y*

Specification of model	A given $Y=y$	Y	Y given $A=i$	A
(i) (U, Y) bivariate normal, A results from partitioning U	Probit regression	Normal	Special skewed distributions of unequal variances (Tables 1, 2)	Arbitrary
(ii) Homogeneous CG-regression chain model with Y as response*	Logistic regression with $\gamma_{ay}^{(l)} = (\mu_1 - \mu_0)/\sigma^2$	Mixture of normals	Normal	Arbitrary
(iii) Homogeneous CG-regression chain model with A as response	Logistic regression with arbitrary γ_{ay}	Normal	Approximately like (i) (2·10)	Arbitrary
(iv) Y normal, $0{\cdot}2 \leqslant \pi_{i\|y}^{A\|Y} \leqslant 0{\cdot}8$ depends linearly on y	Linear-in-probabilities regression	Normal	Approximately normal (2·12)	Requires $0{\cdot}2 \leqslant \pi_i^A \leqslant 0{\cdot}8$
(v) $Y\|(A=i)$ normal $0{\cdot}2 \leqslant \pi_{i\|y}^{A\|Y} \leqslant 0{\cdot}8$ depends linearly on y	Linear-in-probabilities regression	Approximately normal (2·11)	Normal	Requires $0{\cdot}2 \leqslant \pi_i^A \leqslant 0{\cdot}8$

* Equivalant to a CG, conditional Gaussian, distribution for A, Y.

3. Trivariate distributions

3·1. *Preliminaries*

We now consider three response variables, at least one discrete, usually binary. There are many ways of specifying the joint distribution and to some extent the most suitable formulation for a particular application depends on the questions to be asked, for example the kinds of conditional independencies under investigation. There are further questions as to the extent that different models can in practice be distinguished empirically and, corresponding to the discussion of § 2·4, the extent to which tests of independence derived from different models are essentially the same.

3·2. *Three binary response variables*

It is necessary to begin by reviewing methods for economical representation of three binary variables

$$\pi_{ijk}^{ABC} = \mathrm{pr}\,(A=i, B=j, C=k) \quad (i, j, k = 0, 1).$$

Some can be derived by specializing the saturated model

$$\pi_{ijk}^{ABC} = H(\mu^{(h)} + \xi_A^{(h)} i^* + \xi_B^{(h)} j^* + \xi_C^{(h)} k^* + \xi_{AB}^{(h)} i^* j^* + \xi_{AC}^{(h)} j^* k^* + \xi_{BC}^{(h)} j^* k^* + \xi_{ABC}^{(h)} i^* j^* k^*), \tag{3·1}$$

where $i^* = 2i - 1$, etc. and so takes values $(-1, 1)$ as i takes values $(0, 1)$; $\mu^{(h)}$ is a normalizing constant and $H(x)$ is a suitable function. The choices $H(x) = e^x$, $H(x) = x$, $H(x) = \Phi(x)$ are the most common, the first having advantages for the expressions of conditional independencies, the second allowing the simple calculation of marginal distributions, and the last being closely related to tetrachoric correlations, having the longest history. The corresponding parameters are denoted by (l) for log linear, are written without superscript in the linear case, and get the superscript (p) for joint probit in the last case.

The representations above arise when the joint distribution of the variables A, B, C is a natural starting point, treating the three variables on an equal footing. If, however, there is an univariate recursive system with A being a response to B, C, and B being a response to C, a different representation suggests itself. In this case we write

$$\begin{gathered}\pi^{A|BC}_{1|jk} = G(\gamma^{(g)}_{a.bc(bc)} + \gamma^{(g)}_{ab.c(bc)} j^* + \gamma^{(g)}_{ac.b(bc)} k^* + \gamma^{(g)}_{a(bc).bc} j^* k^*), \\ \pi^{B|C}_{1|k} = G(\gamma^{(g)}_{b.c} + \gamma^{(g)}_{bc} k^*), \quad \pi^{C}_{1} = G(\gamma^{(g)}_{c}),\end{gathered} \tag{3·2}$$

where, as before, $j^* = 2j - 1$, $k^* = 2k - 1$ and $G(x)$ is a suitable function. The choices $G(x) = L(x)$, $G(x) = x$ and $G(x) = \Phi(x)$ correspond most directly to those of $H(x)$ discussed for (3·1). The corresponding parameters are denoted by (l) for logit regression, are written without a superscript in the linear regression case and have superscript (p) for probit regression, in the last case.

If $G(x) = L(x)$, the equations in (3·2) are logit regressions with discrete explanatory variables. The parameters in such a regression relate in a simple way to the parameters in a log linear model for the corresponding joint probabilities, since from $G(x) = L(x)$ we get e.g.

$$\log(\pi^{A|BC}_{1|jk} / \pi^{A|BC}_{0|jk}) = \log \pi^{ABC}_{1jk} - \log \pi^{ABC}_{0jk}.$$

Conditional independencies correspond to vanishing logit regression coefficients, for instance $A \perp B \mid C$ is expressed by $0 = \gamma^{(l)}_{a(bc).bc} = \gamma^{(l)}_{ab.c(bc)}$ or $B \perp C$ is expressed by $0 = \gamma^{(l)}_{bc}$. The question of when the same independence structure results from restrictions on systems of recursive logit regressions like (3·2) or more complex ones and from restrictions on a corresponding log linear model has been answered by Wermuth & Lauritzen (1983).

If $G(x) = \Phi(x)$ then the equations in (3·2) are probit regressions with discrete explanatory variables. The parameters in these probit regressions do not relate in a simple way to the parameters in probits for the corresponding joint probabilities except in the case of median dichotomized bivariate normal variates. Note that for a trivariate normal distribution the joint probit model of (3·1) and the corresponding system of recursive probit regressions (3·2) are not equivalent even in the case of median-dichotomizing all three variables. Similarly, the parameters of the linear-in-probabilities regressions, that is $G(x) = x$, do not in general connect simply to the parameters of a linear model for the corresponding joint probabilities; see (A2·14). However, they mimic relations connecting total with partial regression coefficients. For instance, we can write in the case of $\gamma_{a(bc).bc} = 0$:

$$E_{A|C}(\pi^{A|C}_{1|k}) = E_{B|C} E_{A|BC}(\pi^{A|BC}_{1|jk}) = \gamma_{a.c} + \gamma_{ac} k^*,$$

where

$$\gamma_{a.c} = \gamma_{a.bc} + \gamma_{ab.c}\gamma_{b.c}, \quad \gamma_{ac} = \gamma_{ac.b} + \gamma_{ab.c}\gamma_{bc}.$$

Further conditional probabilities such as $\pi^{B|AC}_{j|ik}$ can of course be calculated but involve ratios of combinations of the original parameters. See Appendix 2 for more detailed discussion.

A data set in which recursive linear-in-probabilities regressions give an appropriate description is taken from Weck (1991, p. 182) for $n = 2026$ German students. The three binary variables ($1 = \text{yes}$, $0 = \text{no}$) are: A, change of field study; B, poor integration in high school classes; C, change of primary school. The counts n_{ijk} are

$$(n_{111}, n_{011}, n_{101}, n_{001}, n_{110}, n_{010}, n_{100}, n_{000}) = (15, 33, 84, 278, 40, 113, 246, 1217).$$

The saturated model of the type (3·2) with $G(x)=x$, as in (A2·12), is

$$\hat{\pi}_{1|jk}^{A|BC} = 0{\cdot}244 + 0{\cdot}043j^* + 0{\cdot}029k^* + 0{\cdot}0003j^*k^*,$$

$$\hat{\pi}_{1|k}^{B|C} = 0{\cdot}106 + 0{\cdot}011k^*, \quad \hat{\pi}_1^C = 0{\cdot}202.$$

It is well reproduced assuming $B \perp C$ that is B independent of C, and no interaction effect of B and C on A, that is by taking $\pi_{ijk}^{ABC} = \pi_{i|jk}^{A|BC}\pi_j^B\pi_k^C$ and least squares estimates

$$\tilde{\pi}_{1|jk}^{A|BC} = 0{\cdot}245 + 0{\cdot}045j^* + 0{\cdot}031k^*, \quad \tilde{\pi}_1^B = 0{\cdot}109, \quad \tilde{\pi}_1^C = 0{\cdot}202.$$

This gives as highest risk to change the field of study $\tilde{\pi}_{1|11}^{A|BC} = 0{\cdot}32$ and as lowest $\tilde{\pi}_{1|00}^{A|BC} = 0{\cdot}17$.

3·3. *Two continuous and one binary variable*

In studying two continuous variables X, Y and one binary variable A, there are four types of independence which may be of interest, exemplified by $X \perp Y \mid A$, $A \perp X \mid Y$, $X \perp Y$, $A \perp X$. We shall consider five families of models.

(i) We have a homogeneous conditional Gaussian distribution model in which the distribution of A is arbitrary and in which the conditional distribution of (X, Y) given $A=i$ is bivariate normal with vector mean $\mu_i = (\mu_x(i), \mu_y(i))$ and covariance matrix Σ.

(ii) Secondly we have a homogeneous conditional Gaussian regression chain model with A as response in which the marginal distribution of (X, Y) is bivariate normal with vector mean $\mu = (\mu_x, \mu_y)$ and covariance matrix Σ, and in which given $X=x$, $Y=y$, A has linear logistic regression

$$\pi_{1|xy}^{A|XY} = L\{\gamma_{a.xy}^{(l)} + \gamma_{ax.y}^{(l)}(x-\mu_x) + \gamma_{ay.x}^{(l)}(y-\mu_y)\}. \tag{3·3}$$

(iii) Thirdly we have a dichotomized normal model in which (U, X, Y) are trivariate normal and in which A is formed by dichotomizing U thus producing a model differing from (ii) only in replacing (3·3) by

$$\pi_{1|xy}^{A|XY} = \Phi\{\gamma_{a.xy}^{(p)} + \gamma_{ax.y}^{(p)}(x-\mu_x) + \gamma_{ay.x}^{(p)}(y-\mu_y)\}. \tag{3·4}$$

(iv) Fourthly we have a homogeneous conditional Gaussian regression chain model with (A, X) as joint responses to Y in which Y is marginally normal, and in which A given $Y=y$ has linear logistic form

$$\pi_{1|y}^{A|Y} = L\{\gamma_{a.y}^{(l)} + \gamma_{ay}^{(l)}(y-\mu_y)\}, \tag{3·5}$$

and in which X given both of $A=i$, $Y=y$ has a normal distribution with parallel regression lines on y at the two levels of A is denoted by

$$E(X \mid A=i, Y=y) = \mu_x(i) + \beta_{xy\,a}(y-\mu_y(i)) = \beta_{x.ay}(i) + \beta_{xy.a}y. \tag{3·6}$$

Thus while Y is marginally normal, X is not unless $\mu_x(1) = \mu_x(0)$, but it is close to normality if the standardized difference in means is small, as in (2·11).

(v) Finally we have a linear representation of probabilities in which, over an inevitably restricted range,

$$\pi^{A|XY}_{i|xy} = \pi^A_i + \gamma_{ax.y}(x-\mu_x)i^* + \gamma_{ay.x}(y-\mu_y)i^*, \tag{3·7}$$

and in which (X, Y) is bivariate normal. Marginalization in this model gives

$$\pi^{A|X}_{i|x} = \pi^A_i + (\gamma_{ax.y} + \beta_{yx}\gamma_{ay.x})(x-\mu_x)i^* = \pi^A_i + \gamma_{ax}(x-\mu_x)i^*,$$

say, where β_{yx} is the linear regression coefficient of Y on X. Note that γ_{ax} is determined by a relation of the same form as used in marginalizing least squares regression coefficients.

Choice between these models is partly an empirical matter, although (ii) and (iii) are known to be distinguishable only from very large amounts of data, and model (v), which is subject to the restriction that the right-hand side lies in $(0, 1)$, is essentially the same as models (ii) and (iii) if X and Y are such that with high probability $\pi^{A|XY}_{i|xy}$ is in the range $(0{\cdot}2, 0{\cdot}8)$. Marginal nonnormality of both or one of Y or X points toward (i) or (iv). Also, if we wish to test a particular conditional independence, certain models will be virtually excluded, because under some models particular hypotheses may arise only in a way that demand independencies additional to the one to be tested, implying that such a model is unsuitable for the required purpose.

The hypothesis $X \perp Y \mid A$ is tested from (i) and (iv) via the vanishing of the regression coefficient, say of Y on X, within the two groups of observations with respectively $A=0, 1$, so that a standard normal theory test is available. On the other hand, under the models (ii), (iii), (v), that is those having A as univariate response, $X \perp Y \mid A$ if and only if either $X \perp (A, Y)$ or $Y \perp (A, X)$, and testing either of these would amount to examining a hypothesis much more stringent than the hypothesis of interest initially.

The hypothesis $A \perp X \mid Y$ can be tested from (ii), (iii), (v) via the vanishing of the $\gamma^{(g)}_{ax.y}$ in the binary regression of A on X, Y. Comparison of the tests under the different models is no longer straightforward, in part because the null hypotheses being tested which allow dependence of A on the second variable Y are not the same under the different models. If we take as the model

$$\pi^{A|XY}_{1|xy} = G(\mu + \gamma_x x + \gamma_y y), \tag{3·8}$$

where we have simplified the notation slightly as compared with (3·2), a component of the log likelihood from independent observations (i_j, x_j, y_j) $(j=1, \ldots, n)$ is

$$\mathscr{L} = \sum [i_j \log G(\mu + \gamma_x x_i + \gamma_y y_i) + (1-i_j)\log\{1 - G(\mu + \gamma_x x_j + \gamma_y y_j)\}],$$

and the score statistic for testing $\gamma_x = 0$ is based on $U_{x0} = (\partial\mathscr{L}/\partial\gamma_x)$ evaluated at $\gamma_x = 0$ and at $\mu = \hat{\mu}_0$, $\gamma_y = \hat{\gamma}_{y0}$, the maximum likelihood estimates of (μ, γ_y) at $\gamma_x = 0$. In fact

$$U_{x0} = \sum \hat{W}_{j0} x_j (i_j - \hat{G}_{j0}),$$

where $\hat{G}_{j0} = G(\hat{\mu}_0 + \hat{\gamma}_{y0} y_j)$ and $\hat{W}_{j0} = W(\hat{\mu}_0 + \hat{\gamma}_{y0} y_j)$ with

$$W(x) = G'(x)/[G(x)\{1-G(x)\}].$$

When $G(x) = L(x)$, the unit logistic function, $W(x) = 1$, whereas, when $G(x) = \Phi(x)$, $1 \leq W(x)/W(0) < 1{\cdot}22$ over the range in which $0{\cdot}1 \leq \Phi^{-1}(x) \leq 0{\cdot}9$ with $1{\cdot}22$ replaced by $1{\cdot}10$ in the narrower range $0{\cdot}2 \leq \Phi^{-1}(x) \leq 0{\cdot}8$. For the linear form $G(x) = x$ the weight function varies more strongly with $1{\cdot}22$ and $1{\cdot}10$ replaced by $2{\cdot}18$ and $1{\cdot}56$ respectively.

The expected information matrix for $(\mu, \gamma_x, \gamma_y)$ evaluated at $(\hat{\mu}_0, 0, \hat{\gamma}_{y0})$ is, on differentiating $\mathcal{L}$ twice with respect to the parameters and taking expectations, equal to

$$\begin{bmatrix} \sum \hat{V}_{j0} & \sum \hat{V}_{j0} x_j & \sum \hat{V}_{j0} y_j \\ . & \sum \hat{V}_{j0} x_j^2 & \sum \hat{V}_{j0} x_j y_j \\ . & . & \sum \hat{V}_{j0} y_j^2 \end{bmatrix}, \tag{3·9}$$

where $\hat{V}_{j0} = V(\hat{\mu}_0 + \hat{\gamma}_{y0} y_j)$ with $V(x) = G'(x) W(x)$ and the null hypothesis variance of V_{x0} is the reciprocal of the $(2, 2)$ element in the inverse of (3·9). In these calculations the $\{x_j, y_j\}$ are regarded as fixed, as would be appropriate in analyzing a given set of observations. For some theoretical purposes, however, we may take expectations over the marginal bivariate normal distribution of (X, Y). If in the fitting we measure x_j and y_j as deviations from their sample means, μ becomes orthogonal to (γ_x, γ_y) and the 2×2 information matrix for the latter is

$$n \begin{bmatrix} E\{V(\mu + \gamma_y Y) X^2\} & E\{V(\mu + \gamma_y Y) XY\} \\ . & E\{V(\mu + \gamma_y Y) Y^2\} \end{bmatrix}.$$

In the not very interesting case where $\gamma_y \simeq 0$, it follows immediately that the test based on U_{x0} is asymptotically independent of the form of the function $G(.)$. More generally, even if strong regression of A on y is present, the tests based on two different $G(.)$'s that give essentially the same fit to the dependence on y are unlikely to differ appreciably. For in (3·9), the $\hat{G}_{j0}$ will not differ much and the weights $\hat{W}_{j0}$ vary in a limited way as noted above. If, however, tests use forms of $G(.)$ that differ notably in their fit to the dependence of A on y, then of course the tests are different, especially if X and Y are strongly correlated. The reason is that the use of an inappropriate function for 'adjusting for' the dependence on y could induce bias in the test of $\gamma_x = 0$, quite apart from questions of efficiency.

Under the homogeneous conditional Gaussian distribution model (i), $A \perp X \mid Y$ requires that the regression lines of X on Y for $A = 0$, 1 coincide.

The hypothesis $X \perp Y$ is directly tested in (ii), (iii) and (v) via a test of independence in the postulated bivariate normal distribution of (X, Y), whereas, under (i) and (iv), the hypothesis is satisfied only if $(A, X) \perp Y$ or $(A, Y) \perp X$.

Finally $A \perp X$ is directly tested in (i) via the equality of the means of X at $A = 0, 1$. The linear form (v) is marginalized by taking expectations conditionally on $X = x$ to give (2·9), so that the required independence $A \perp X$ is directly tested from the regression of A on X or of X on A giving tests which, as discussed in § 2·4, are asymptotically equivalent to that from (i).

The same test can also be derived from the probit model (iii). Because of the close numerical equivalence of probit and logistic forms, the same test will also be effective under the logistic model in (iv) and in (ii) although in the latter case only as a result of a mathematical approximation; compare (2·10). Thus whenever a hypothesis can be tested via a number of different models, the resulting test is usually approximately the same regardless of the model, confirming the more detailed analysis of § 2·4 for two variables.

In summary, therefore, the various models give at least roughly equivalent tests for the various independence hypotheses, provided the model considered does not make the hypothesis in question collapse into a stronger form of independence; see Table 5. Typically, such problems will not occur if only hypotheses are to be tested which

Table 5. *Null hypotheses under different distributional assumptions for A, X, Y*

Specification of model	Hypothesis			
(§ 3·3)	$A \perp X \mid Y$	$X \perp Y \mid A$	$A \perp Y$	$X \perp Y$
(i) Homogeneous CG-regression chain model with (X, Y) as response to A*	$\beta_{x.ay}(1) = \beta_{x.ay}(0)$ in (3·6)	$\beta_{xy.a} = 0$ in (3·6)	$\mu_y(1) = \mu_y(0)$	Only if either $(A, X) \perp Y$ or $(A, Y) \perp X$
(ii) Homogeneous CG-regression chain model with A as response to (X, Y)	$\gamma^{(l)}_{ax.y} = 0$ in (3·3)	Only if either $X \perp (A, Y)$ or $Y \perp (A, X)$	Approximately like (iii)	$\rho_{xv} = 0$
(iii) Trivariate normal for (X, Y, U) with U dichotomized to form A	$\gamma^{(p)}_{ax.y} = 0$ in (3·4)	Same as (ii)	$\gamma^{(p)}_{ay} = 0$	Same as (ii)
(iv) Homogeneous CG-regression chain model with (X, A) as response to Y†	Same as (i)	Same as (i)	$\gamma^{(l)}_{ay} = 0$ in (3·5)	Same as (i)

* Equivalent to a CG, conditional Gaussian, distribution for A, X, Y.

† Marginal independence $A \perp X$ in this model and in a corresponding model having logistic dependence of A on Y replaced by a probit dependence requires $(A, Y) \perp X$ or $A \perp (X, Y)$, while $A \perp Y \mid X$ requires $\gamma^{(l)}_{ay} = \beta_{xy.a}\{\mu_x(1) - \mu_x(0)\}/\sigma^2_{x.ya}$.

correspond to the ordering, the conditioning of variables, implied by the dependence chain used to specify the model, i.e. whenever the dependence chain and where the independencies to be tested result from substantive considerations (Wermuth & Lauritzen, 1990).

3·4. *Two binary and one continuous variable*

We now carry out a broadly parallel analysis to §§ 3·2, 3·3 when there are two binary variables A, B and one continuous variable, X. The independency relations of interest are exemplified by $A \perp B \mid X$, $A \perp X \mid B$, $A \perp B$ and $A \perp X$. We consider five families of models.

(i) We have a homogeneous conditional Gaussian distribution model in which the distribution of (A, B) is arbitrary, with probabilities π^{AB}_{ij}, and in which X given $A = i$, $B = j$ is normally distributed with mean μ_{ij} and variance σ^2.

(ii) We have a homogeneous conditional Gaussian regression chain model with (A, B) as responses to X, with X having a marginal normal distribution of mean μ and variance σ^2, and A, B having a joint log linear model given X written conveniently in the form

$$\pi^{AB|X}_{ij|x} \propto \exp\{\mu^{(l)} + \xi^{(l)}_{A} i^* + \xi^{(l)}_{B} j^* + \xi^{(l)}_{AB} i^* j^* + \xi^{(l)}_{AX} i^*(x - \mu_x) + \xi^{(l)}_{BX} j^*(x - \mu_x) + \xi^{(l)}_{ABX} i^* j^*(x - \mu_x)\}. \quad (3{\cdot}10)$$

(iii) We have a homogeneous conditional Gaussian regression chain model with A as a response to (B, X), that is with X given $B = j$ normal with mean μ_{xj} and variance σ^2_{xb} and with A given $X = x$, $B = j$ logistic

$$\pi^{A|XB}_{1|xj} = L\{\gamma^{(l)}_{a.xb(xb)} + \gamma^{(l)}_{ax.b(xb)}(x - \mu_{xj}) + \gamma^{(l)}_{ab.x(xb)} j^* + \gamma^{(l)}_{a(xb).bx}(x - \mu_{xj}) j^*\}. \quad (3{\cdot}11)$$

(iv) A number of models can be formed from underlying multivariate normal distributions, one by replacing $L(x)$ in (3·11) by $\Phi^{-1}(x)$, this corresponding to bivariate normal distributions for (U, X) given $B=j$, and another corresponding to a trivariate normal distribution for (U, V, X), the first two variables being dichotomized to form (A, B).

(v) Finally, there are representations linear in the probabilities which can be written as

$$\pi^{AB|X}_{ij|x} = \tfrac{1}{4}\{1+\xi_A i^* + \xi_B j^* + \xi_{AB} i^* j^* + \xi_{AX} i^*(x-\mu_x) + \xi_{BX} j^*(x-\mu_x) + \xi_{ABX} i^* j^*(x-\mu_x)\}, \qquad (3{\cdot}12)$$

or as (3·11) with L and superscript (l) deleted. They have advantages of simple marginalization and direct interpretation but the disadvantages of inevitably constrained parameters and give only indirect representations of some conditional independencies.

Discussion of the tests of various kinds of independency parallels that of § 3·3. Programmed algorithms for estimation in conditional Gaussian distributions are due to Edwards (1990).

Thus, for example, $A \perp B \mid X$ is tested quite directly in the conditional Gaussian regression chain models (ii) and (iii), for example in (iii) by testing $\gamma^{(l)}_{ab.x(xb)} = \gamma^{(l)}_{a(xb).bx} = 0$ in the linear logistic regression of A on X, B and XB associated with (3·11), and in (ii) the same test is appropriate, because (3·10) implies a relation of the type (3·11) for the conditional distribution of A given X and B. On the other hand, under the conditional Gaussian distribution model (i), $A \perp B \mid X$ requires both that the normal means μ_{ij} have an additive structure $\mu_{ij} = \mu + \xi_A i^* + \xi_B j^*$ and that the marginal odds ratio takes a special value

$$(\pi^{AB}_{11}\pi^{AB}_{00})/(\pi^{AB}_{10}\pi^{AB}_{01}) = \exp(4\xi_A\xi_B/\sigma^2)$$

(Wermuth, 1989). A likelihood ratio test can be set up for this hypothesis, but the precise relation between it and the tests associated with models (i) and (iii) is unclear. An exception is the case in which the linear-in-probability regressions approximate both of $\pi^{A|X}_{i|x}$ and $\pi^{B|X}_{j|x}$ well. Then the results of (2·11), (2·12) imply that the models are virtually indistinguishable under the hypothesis $A \perp B \mid X$, no matter whether the distributional assumptions (i), (ii) or (iii) hold.

For $A \perp X \mid B$, the conditional Gaussian distribution model (i) is immediately applicable via a test of no interaction in the two-way analysis of the cell means. Under a conditional Gaussian regression chain model with discrete responses, $A \perp X \mid B$ requires $\xi^{(l)}_{AX} = \xi^{(l)}_{ABX} = 0$ for (3·10) and $\gamma^{(l)}_{ab.x(xb)} = \gamma^{(l)}_{a(xb).bx} = 0$ for (3·11) and the two tests are equivalent since, as mentioned before, (3·10) implies a relation of the type (3·11).

3·5. *Concluding remarks*

Most of the models discussed above treat the three involved variables asymmetrically. This can lead to particularly simple and appealing interpretations if single variables are responses and it is especially important when it can be given a substantive interpretation. Nevertheless all the models are to be regarded as specifying the joint distribution of three random variables involved. A particular conditional hypothesis, which can be directly specified and tested as independence of a response from one of the explanatory variables, may not be satisfied in some joint distribution unless a stronger independence holds. One example is $X \perp Y \mid A$ which is a common hypothesis in a linear regression of X or Y on the remaining variables, but which cannot hold if A arises as a dichotomized variable from U, where X, Y, U have a joint normal distribution. A similar but not completely analogous case is $A \perp B \mid Y$, which is a common hypothesis in a probit regression of A or B on the remaining variables, but an unstable hypothesis in a conditional Gaussian distribution of Y, A, B, that is, even though the hypothesis can be satisfied by some expected counts, the sample size has to be very large to distinguish it from one of the stronger hypotheses $A \perp (B, Y)$ or $B \perp (A, Y)$.

If in addition there is for each individual a vector z of explanatory variables which can be treated as if fixed, the addition to the models of a term for linear dependence on z is in most cases as discussed in § 2·5.

An important qualitative conclusion is that there is a variety of models for representing this kind of data and that to a considerable extent tests of conditional independence do not depend strongly on model choice. This allows some flexibility of choice in selecting models that are convenient for substantive interpretation and for probability calculations.

Broadly similar results apply to situations with more than three variables, but conceptually new problems may also arise if distributions of four or more variables are studied. For instance, some models will contain so-called nondecomposable independence hypotheses, i.e. independencies which cannot be conveniently specified by zero restrictions on individual parameters of recursive systems such as (3·2), but which involve associated joint responses instead. As a consequence not only the interpretation can be more difficult, but the available estimation procedures for sequences of univariate recursive regressions have to be extended to obtain estimates. Noniterative approximations may be utilized in some situations (Cox & Wermuth, 1990, 1991), but typically iterative algorithms are required to obtain maximum likelihood estimates under such more complex models.

Acknowledgement

We are grateful to the British German Academic Research Collaboration Programme for supporting our work.

Appendix 1

Least squares analysis of linear representations of probabilities

Let $Y_1, \ldots, Y_n$ be independent binary random variables with

$$E(Y_j) = \mathrm{pr}\,(Y_j = 1) = \pi_j, \quad \mathrm{var}\,(Y_j) = \pi_j(1 - \pi_j) = \bar{\nu}(1 + \delta_j),$$

where $\bar{\nu} = \mathrm{ave}\,\{\pi_j(1 - \pi_j)\}$ and $\Sigma\, \delta_j = 0$. Consider a linear representation of the probabilities, $\pi_j = x_j/\beta$, where x_j is a $1 \times p$ and β is a $p \times 1$ vector of parameters. If Y is the $n \times 1$ vector of the y_j, then $E(Y) = x\beta$, where x is $n \times p$ and assumed to be of full rank.

We suppose that the $\{\pi_j\}$ are confined to a central range such as (0·2, 0·8), so that the constraints $0 \leq \pi_j \leq 1$ can be ignored. Also the $\{\delta_j\}$ are then small.

The ordinary least squares estimates $\hat{\beta}_{\mathrm{LS}} = (x^{\mathrm{T}}x)^{-1}x^{\mathrm{T}}Y$ have covariance matrix

$$\mathrm{cov}\,(\hat{\beta}_{\mathrm{LS}}) = \bar{\nu}(x^{\mathrm{T}}x)^{-1}\{I + (x^{\mathrm{T}}\Delta x)(x^{\mathrm{T}}x)^{-1})\}, \qquad (\mathrm{A}1{\cdot}1)$$

where $\Delta = \mathrm{diag}\,(\delta_1, \ldots, \delta_n)$. Direct calculation shows that the expected value of the residual mean square is

$$\bar{\nu}[1 - (n - p)^{-1}\,\mathrm{tr}\,\{(x^{\mathrm{T}}\Delta x)(x^{\mathrm{T}}x)^{-1}\}], \qquad (\mathrm{A}1{\cdot}2)$$

so that some adjustment is in principle desirable in attaching standard errors to the components of $\hat{\beta}_{\mathrm{LS}}$.

The log likelihood function is

$$L(\beta)=\sum_j \{y_j \log (x_j\beta)+(1-y_j)\log (1-x_j\beta)\}$$

and, on differentiating twice with respect to β and taking expectations, it follows that the Fisher information matrix for β is

$$\bar{\nu}^{-1}x^{\mathrm{T}}\operatorname{diag}\{(1+\delta_j)^{-1}\}x=\bar{\nu}^{-1}x^{\mathrm{T}}x\{I-(x^{\mathrm{T}}x)^{-1}(x^{\mathrm{T}}\Delta x)+(x^{\mathrm{T}}x)^{-1}(x^{\mathrm{T}}\Delta^2x)+O(\Delta^3)\}. \quad (\mathrm{A1{\cdot}3})$$

Thus the asymptotic covariance matrix of $\hat{\beta}_{\mathrm{ML}}$, the maximum likelihood estimate, is on inversion

$$\bar{\nu}(x^{\mathrm{T}}x)^{-1}[I+(x^{\mathrm{T}}\Delta x)(x^{\mathrm{T}}x)^{-1}-x^{\mathrm{T}}\Delta^{\mathrm{T}}\{I-x(x^{\mathrm{T}}x)^{-1}x^{\mathrm{T}}\}\Delta x(x^{\mathrm{T}}x)^{-1}+O(\Delta^3)]. \quad (\mathrm{A1{\cdot}4})$$

Comparison with (A1·1) shows that the inflation of variance by using $\hat{\beta}_{\mathrm{LS}}$ rather than $\hat{\beta}_{\mathrm{ML}}$ is of order Δ^2. This could have been anticipated from the identity between maximum likelihood estimation and weighted least squares with appropriately iterated weights and the known insensitivity of weighted least squares to perturbations in the weights.

As a rather extreme case consider the model with

$$E(Y_j)=\beta_0 \quad (j=1,\ldots,n_1), \quad E(Y_j)=\beta_0-\beta_1 \quad (j=n_1+1,\ldots,n_1+\tfrac{1}{2}n_2),$$
$$E(Y_j)=\beta_0+\beta_1 \quad (j=n_1+\tfrac{1}{2}n_2+1,\ldots,n_1+n_2),$$

with in fact $\beta_0=\frac{1}{2}$, $\beta_1=0{\cdot}3$, so that maximal changes of variance are encountered. Then

$$\operatorname{var}(\hat{\beta}_{0,\mathrm{LS}})/\operatorname{var}(\hat{\beta}_{0,\mathrm{ML}})=(n_1+2{\cdot}2025n_1n_2+n_2^2)/(n_1+n_2)^2<1{\cdot}051.$$

It thus seems likely that the loss of efficiency is typically less than, and often much less than, 5%.

These arguments can be extended to any simple exponential family problem involving independent observations in which the mean parameter is specified by a linear model.

Appendix 2

Models linear in probabilities

In this appendix we develop further some aspects of models linear in probabilities as set out in § 3. The advantages of such models are that parameters are directly interpreted via differences or contrasts of probabilities, that simple marginalization is available by addition of probabilities and that fitting by ordinary least squares is often highly efficient. The disadvantages are that independence is a multiplicative rather than an additive concept and that constraints on the parameters are unavoidable unless the probabilities are restricted to a central range.

We discuss separately the linear models based on the joint distribution (3·1) and the linear models formulated recursively (3·2).

First note that it is possible to augment (3·1) by terms depending on one or more quantitative variables as in (3·12) and indeed we can include terms in $(X-\mu_x)^2$ if desired. Taking expectations over X in (3·12) leads back to the model for π_{ij}^{AB}.

The relative clumsiness of the linear representation for π_{ij}^{AB} in dealing with independence and conditional distributions is shown by formulae like

$$\pi_i^A=\tfrac{1}{2}(1+\xi_A i^*), \quad \pi_j^B=\tfrac{1}{2}(1+\xi_B j^*),$$
$$\pi^A{}_{i|j}^{\,|B}=\tfrac{1}{4}(1+\xi_A i^*+\xi_B j^*+\xi_{AB}i^*j^*)/\{\tfrac{1}{2}(1+\xi_B j^*)\}. \quad (\mathrm{A2{\cdot}1})$$

Thus, A and B are independent if and only if $\xi_{AB}=\xi_A\xi_B$ so that independence can be assessed via the nonlinear combination

$$\eta_{AB}=\xi_{AB}-\xi_A\xi_B. \quad (\mathrm{A2{\cdot}2})$$

This is equivalent to $\xi_{AB}=0$ if and only if at least one of A, B is equally likely to take values zero, one. Introduction of the categories i^*, etc. taking values $(1,-1)$ instead of i taking values

(0, 1) is not essential but does symmetrize the formulae. Note that, if I^* is the random variable corresponding to A, then $E(I^*) = \xi_A$, var $(I^*) = 1 - \xi_A^2$.

Similarly for three variables A, B, C starting from

$$\pi_{ijk}^{ABC} = \tfrac{1}{8}(1 + \xi_A i^* + \xi_B j^* + \xi_C k^* + \xi_{AB} i^* j^* + \xi_{AC} i^* k^* + \xi_{BC} j^* k^* + \xi_{ABC} i^* j^* k^*), \qquad \text{(A2·3)}$$

conditional independence $A \perp B \mid C$ involves two conditions

$$\pi_{111}^{ABC} / \pi_{11}^{BC} = \pi_{110}^{ABC} / \pi_{10}^{BC}, \quad \pi_{101}^{ABC} / \pi_{01}^{BC} = \pi_{100}^{ABC} / \pi_{01}^{BC},$$

leading to

$$\eta_{AC}\eta_{BC} = \eta_{AB}(1 - \xi_C^2), \quad \xi_{ABC} - \xi_A \xi_B \xi_C = \xi_A \eta_{BC} + \xi_B \eta_{AC} + \xi_C \eta_{AB}. \qquad \text{(A2·4)}$$

Under complete independence $A \perp B \perp C$ we have that

$$\eta_{AB} = \eta_{AC} = \eta_{BC} = 0, \quad \xi_{ABC} = \xi_A \xi_B \xi_C, \qquad \text{(A2·5)}$$

suggesting that it may sometimes be convenient to define

$$\eta_{ABC} = \xi_{ABC} - \xi_A \xi_B \xi_C. \qquad \text{(A2·6)}$$

Note that when the marginal probabilities are equal or close to $\frac{1}{2}$ as when binary variables are produced by median dichotomizing of continuous variables, then $\xi_A = \xi_B = \xi_C = 0$ and equations (A2·4)-(A2·6) simplify appreciably.

Conditional distributions can be written down in forms exemplified by

$$\pi_{i|jk}^{A|BC} = \pi_{ijk}^{ABC} / \{\tfrac{1}{4}(1 + \xi_B j^* + \xi_C k^* + \xi_{BC} j^* k^*)\}, \quad \pi_{ij|k}^{AB|C} = \pi_{ijk}^{ABC} / \{\tfrac{1}{2}(1 + \xi_C k^*)\}. \qquad \text{(A2·7)}$$

A second set of useful linear representations can be obtained when the variables are ordered to have A as response to (B, C) and B as response to C, that is in such a way that it is sensible to build up the joint distribution from the marginal distribution of C, the conditional distribution of B given C and the conditional distribution of A given B and C. We write, in a notation chosen to stress the relation with least squares regression formulae,

$$\pi_k^C = \tfrac{1}{2}(1 + \xi_C k^*), \quad \pi_{j|k}^{B|C} = \tfrac{1}{2}(1 + \gamma_{b.c} j^* + \gamma_{bc} j^* k^*). \qquad \text{(A2·8)}$$

Thus the marginal distribution of B has

$$\pi_j^B = \tfrac{1}{2}(1 + \gamma_{b.c} j^* + \gamma_{bc} \xi_C j^*) = \tfrac{1}{2}(1 + \xi_B j^*),$$

where $\xi_B = \gamma_{b.c} + \gamma_{bc}\xi_C$. Note also that cov $(J^*, K^*) = \gamma_{bc}(1 - \xi_C^2)$ so that the regression coefficient of J^* on K^* is γ_{bc}. This can be found, for example, from the joint distribution of B and C obtained by multiplying the two equations (A2·8) using $k^{*2} = 1$, that is

$$\pi_{jk}^{BC} = \tfrac{1}{4}\{1 + (\gamma_{b.c} + \gamma_{bc}) j^* + \xi_C k^* + \gamma_{bc} j^* k^*\} = \tfrac{1}{4}(1 + \xi_B j^* + \xi_C k^* + \xi_{BC} j^* k^*), \qquad \text{(A2·9)}$$

where

$$\xi_B = \gamma_{b.c} + \gamma_{bc}\xi_C, \quad \xi_{BC} = \gamma_{bc}, \qquad \text{(A2·10)}$$

establishing a connection with the direct specification via the joint distribution. Also

$$E(J^* \mid K^* = k^*) = \gamma_{b.c} + \gamma_{bc} k^*. \qquad \text{(A2·11)}$$

Next write

$$\pi_{i|jk}^{A|BC} = \tfrac{1}{2}\{1 + \gamma_{a.bc} i^* + \gamma_{ab.c(bc)} i^* j^* + \gamma_{ac.b(bc)} i^* k^* + \gamma_{a(bc).bc} i^* j^* k^*\}. \qquad \text{(A2·12)}$$

On taking expectations over the levels of B given C, using (A2·11), we have that

$$\begin{aligned}\pi_{i|k}^{A|C} &= \tfrac{1}{2}\{1 + (\gamma_{a.bc} + \gamma_{ab.c(bc)}\gamma_{b.c} + \gamma_{a(bc).bc}\gamma_{bc}) i^* + (\gamma_{ac\,b(bc)} + \gamma_{ab.c(bc)}\gamma_{bc} + \gamma_{a(bc).bc}\gamma_{b.c}) i^* k^*\} \\ &= \tfrac{1}{2}(1 + \gamma_{a.c} i^* + \gamma_{ac} i^* k^*), \qquad \text{(A2·13)}\end{aligned}$$

say, and again the relations linking the coefficients in (A2·13) with those in (A2·8), (A2·9) and (A2·12) are ordinary regression ones.

Finally, the joint distribution of A, B, C is obtained by multiplying (A2·9) and (A2·12) to form (A2·3) with

$$\begin{aligned}\xi_A &= \gamma_{a.bc} + \xi_B\gamma_{ab.c(bc)} + \xi_C\gamma_{ac.b(bc)} + \xi_{BC}\gamma_{a(bc).bc},\\ \xi_{AB} &= \gamma_{ab.c(bc)} + \xi_B\gamma_{a.bc} + \xi_C\gamma_{a(bc).bc} + \xi_{BC}\gamma_{ac.b(bc)},\\ \xi_{AC} &= \gamma_{ac.b(bc)} + \xi_C\gamma_{a.bc} + \xi_B\gamma_{a(bc).bc} + \xi_{BC}\gamma_{ab.c(bc)},\\ \xi_{ABC} &= \gamma_{a(bc)\ bc} + \xi_B\gamma_{ac.b(bc)} + \xi_C\gamma_{ab.c(bc)} + \xi_{BC}\gamma_{a.bc},\end{aligned} \tag{A2·14}$$

with the ξ's on the right-hand side having already been defined. These simplify considerably if there is no three-factor interaction, leading, for instance, to

$$\gamma_{a.bc} = \xi_A - \gamma_{ab.c}\xi_B - \gamma_{ac.b}\xi_C, \tag{A2·15}$$

and if, in addition, the marginal probabilities are all equal to $\frac{1}{2}$ we have

$$\gamma_{ab.c} = \frac{\xi_{AB} - \xi_{AC}\xi_{BC}}{1-\xi_{BC}^2}, \quad \gamma_{ac.b} = \frac{\xi_{AC} - \xi_{AB}\xi_{BC}}{1-\xi_{BC}^2}. \tag{A2·16}$$

If, however, conditional relations are required in which the order with A as response to B, C and B as response to C is not preserved, then in the present parameterization the linear structure is lost and we return to the form (A2·1).

REFERENCES

ARANDA-ORDAZ, F. J. (1981). On two families of transformations to additivity for binary response data. *Biometrika* **68**, 357–63.

BIRCH, M. W. (1963). Maximum-likelihood in three-way contingency tables. *J. R. Statist. Soc.*, Suppl. **5**, 171–6.

COX, D. R. (1958). The regression analysis of binary sequences (with discussion). *J. R. Statist. Soc.* B **20**, 215–42.

COX, D. R. (1966). Some procedures connected with the logistic response curve. In *Research Papers in Statistics, Essays in Honour of J. Neyman's 70th birthday*, Ed. F. N. David, pp. 55–71. London: Wiley.

COX, D. R. & HINKLEY, D. V. (1974). *Theoretical Statistics.* London: Chapman and Hall.

COX, D. R. & SNELL, E. J. (1989). *Analysis of Binary Data*, 2nd ed. London: Chapman and Hall.

COX, D. R. & WERMUTH, N. (1990). An approximation to maximum likelihood estimates in reduced models. *Biometrika* **77**, 747–61.

COX, D. R. & WERMUTH, N. (1991). A simple approximation for bivariate and trivariate normal integrals. *Int. Statist. Rev.* **59**, 263–9.

EDWARDS, D. (1990). Hierarchical mixed interaction models (with discussion). *J. R. Statist. Soc.* B **52**, 3–20.

FINNEY, D. J. (1952). *Probit Analysis*, 2nd ed. Cambridge University Press.

LAURITZEN, S. L. & WERMUTH, N. (1989). Graphical models for associations between variables, some of which are qualitative and some quantitative. *Ann. Statist.* **17**, 31–57.

MARITZ, J. S. (1953). Estimation of the correlation coefficient in the case of a bivariate normal population when one of the variables is dichotomized. *Psychometrika* **18**, 97–110.

MCCULLAGH, P. & NELDER, J. A. (1989). *Generalized Linear Models*, 2nd ed. London: Chapman and Hall.

PERSON, K. (1901). Mathematical contributions to the theory of evolution—VII. On the correlation of characters not quantitatively measurable. *Phil. Trans. R. Soc. Lond.* A **195**, 1–47.

PEARSON, K. (1903). Mathematical contributions to the theory of evolution—XI. On the influence of natural selection on the variability and correlation of organs. *Phil. Trans. R. Soc. Lond.* A **200**, 1–66.

PRINCE, J. & TATE, R. F. (1966). Accuracy of maximum likelihood estimates of correlation for a biserial model. *Psychometrika* **31**, 85–92.

SOPER, H. E. (1915). On the probable error for the bi-serial expression for the correlation coefficient. *Biometrika* **10**, 384–90.

TALLIS, G. M. (1961). The moment generating function of the truncated multinormal distribution. *J. R. Statist. Soc.* B **23**, 233–9.

TATE, R. F. (1955). The theory of correlation between two continuous variables when one variable is dichotomized. *Biometrika* **42**, 205–16.

WECK, M. P. (1991). *Der Studienfachwechsel. Eine Längstschnittanalyse der Interaktionsstruktur von Bedinungen des Studienverlaufs.* Frankfurt: Lang.

WERMUTH, N. (1989). Moderating effects of subgroups in linear models. *Biometrika* **76**, 81–92.

WERMUTH, N. & LAURITZEN, S. L. (1983). Graphical and recursive models for contingency tables. *Biometrika* **70**, 537-52.

WERMUTH, N. & LAURITZEN, S. L. (1990). On substantive research hypotheses, conditional independence graphs and graphical chain models (with discussion). *J. R. Statist. Soc.* B **52**, 21-72.

[*Received March* 1991. *Revised October* 1991]