Shih, W. J. (1992) On informative and random dropouts in longitudinal studies. *Biometrics*, **48**, 971–972.
Shih, W. J., Quan, H. and Chang, M. N. (1993) Estimation of the mean when data contain non-ignorable missing values from a random effects model. *Statist. Probab. Lett.*, **19**, in the press.
Stefanski, L. A. and Carroll, R. A. (1985) Covariate measurement error logistic regression. *Ann. Statist.*, **13**, 1335–1351.
Tu, X. M., Meng, X. L. and Pagano, M. (1993) The AIDS epidemic: estimating survival after AIDS diagnosis from surveillance data. *J. Am. Statist. Ass.*, **88**, 26–36.
Tukey, J. W. (1986) Comments on "Alternative methods for solving the problem of selection bias" by J. D. Heckman and R. Robb. In *Drawing Inferences from Self Selected Samples* (ed. H. Wainer), pp. 108–110. New York: Springer.
Verbyla, A. P. and Cullis, B. R. (1990) Modelling in repeated measures experiments. *Appl. Statist.*, **39**, 341–356.
——— (1992) The analysis of multistratum and spatially correlated repeated measures data. *Biometrics*, **48**, 1015–1032.
Wu, M. C. and Bailey, K. (1988) Analyzing changes in the presence of informative right censoring caused by death and withdrawal. *Statist. Med.*, **7**, 337–346.
——— (1989) Estimation and comparison of changes in the presence of informative right censoring: conditional linear model. *Biometrics*, **45**, 939–955.
Wu, M. C. and Carroll, R. J. (1988) Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics*, **44**, 175–188.
Wu, M. C., Hunsberger, S. and Zucker, D. (1993) Testing for changes in the presence of censoring: parametric and nonparametric methods. *Statist. Med.*, to be published.

*Appl. Statist.* (1994)
**43**, *No.* 1, *pp.* 95–108

# Logistic Regression for Correlated Binary Data

By S. LE CESSIE† and J. C. VAN HOUWELINGEN

*University of Leiden, The Netherlands*

[Received August 1991. Final revision August 1992]

SUMMARY

The modelling of correlated binary outcomes, in such a way that the marginal response probabilities are still logistic, is considered. Different association measures for the dependence between correlated observations are discussed. For paired correlated data the full likelihood can be evaluated; for an arbitrary number of correlated observations a pseudolikelihood approach to obtain parameter estimates is proposed. The results are illustrated on data from a Dutch follow-up study on preterm infants.

*Keywords*: Correlated outcomes; Logistic regression; Odds ratio; Pseudolikelihood; Tetrachoric correlation

## 1. Introduction

In this paper we discuss the problems which arise when modelling binary data with correlated outcomes. As an example we consider data from a Dutch follow-up study on preterm infants, named 'POPS' (Verloove and Verwey, 1988)—the 'Project on preterm and small for gestational age infants in the Netherlands'. In this study, data were collected on 1338 infants, born in 1983 in the Netherlands, with a gestational age of less than 32 completed weeks and/or a birthweight of less than 1500 g. Neonatal mortality and morbidity (i.e. mortality and morbidity within 28 days) were studied and the surviving infants were re-examined after 2 years and 5 years. A problem when modelling the various binary outcome variables of these data with logistic regression is that the assumption of independent observations does not hold. Many preterm infants result from multiple birth and these infants are likely to respond similarly. For example there are 107 complete pairs of twins in the cohort.

The purpose of this paper is to model the data by taking the dependence between observations into account. A discussion of the various approaches to model correlated binary observations can be found in Prentice (1988), Zeger *et al.* (1988), Neuhaus *et al.* (1991) and in Liang *et al.* (1992). In this paper we consider a 'population-averaged' approach in which the data are modelled such that the marginal outcome probabilities are still logistic. In this way the parameters reflect the effect of the various covariates on the average response probability, instead of the

†*Address for correspondence*: Department of Medical Statistics, University of Leiden, PO Box 9604, 2300 RC Leiden, The Netherlands.

 0035-9254/94/43095

effects among infants of one twin or triplet, and can be interpreted as population-averaged logarithms of adjusted odds ratios.

Ignoring the dependence between observations could lead to incorrect estimates of the standard errors. Furthermore if the outcomes are positively correlated the probability that both twins die is larger than the product of the marginal probabilities and modelling the dependence is needed to obtain correct estimates of the joint probabilities.

In this paper different ways to quantify the dependence are discussed. We start with the simplified problem of paired binary data and apply the results on a subset of the POPS data consisting of the 107 pairs of twins, of which both infants are in the study. This means that both infants are live-born and fulfil the intake criteria.

Generalization to the data of all twins can be made straightforwardly because the bivariate model can just as easily handle a combination of blocks of size 1 and 2. Problems arise when the blocks of correlated observations may have arbitrary sizes. In Section 4 we study the general situation. We use a pseudolikelihood approach to obtain estimates for the pairwise dependence and the parameters of the marginal probabilities and apply these results to POPS data, concerning all infants in the study.

## 2. Bivariate Binary Data

We consider the following type of data, where we have paired outcomes $(Y_1, Y_2)$ and $m$ pairs of observations. We assume that the observations within pairs are correlated but that observations from different pairs are independent. The marginal outcome probabilities are denoted by $p_1$ and $p_2$. The situation is sketched in Table 1. The marginal probabilities are logistic and depend on the same parameter column vector $\beta$. This means that for an observation with covariate vector $x$

$$\Pr(Y=1) = \frac{\exp(x\beta)}{1+\exp(x\beta)}.$$

Extensions to the situation where $p_1$ and $p_2$ depend on different parameters are easy to make.

We assume that the dependence between two observations can be quantified with an extra parameter $\theta$ which we do not specify for the moment. The log-likelihood function for these data is

$$l(\beta,\theta) = \sum_{i=1}^{m} (Y_{11i}\log p_{11i} + Y_{10i}\log p_{10i} + Y_{01i}\log p_{01i} + Y_{00i}\log p_{00i}), \qquad (2.1)$$

TABLE 1
*Different outcomes with probability of occurrence*

| | $Y_2=1$ | $Y_2=0$ | |
|---|---|---|---|
| $Y_1=1$<br>$Y_1=0$ | $p_{11}$<br>$p_{01}$ | $p_{10}$<br>$p_{00}$ | $p_1$<br>$1-p_1$ |
| | $p_2$ | $1-p_2$ | 1 |

where $Y_{jki}=1$ if $Y_{1i}=j$ and $Y_{2i}=k$ for pair $i$, and $Y_{jki}=0$ otherwise, $p_{jki}=\Pr(Y_{jki}=1)$ and where summing is done over all pairs of observations. For each pair $i$, the sum of the probabilities $\Sigma_{j=0,k=0}^{1,1} p_{jki}=1$.

Differentiation with respect to $\beta$ and $\theta$ yields

$$\frac{\partial l(\beta,\theta)}{\partial \beta} = \sum\left(\frac{Y_{11i}}{p_{11i}}\frac{\partial p_{11i}}{\partial \beta}+\frac{Y_{10i}}{p_{10i}}\frac{\partial p_{10i}}{\partial \beta}+\frac{Y_{01i}}{p_{01i}}\frac{\partial p_{01i}}{\partial \beta}+\frac{Y_{00i}}{p_{00i}}\frac{\partial p_{00i}}{\partial \beta}\right),$$
$$\frac{\partial l(\beta,\theta)}{\partial \theta} = \sum\left(\frac{Y_{11i}}{p_{11i}}-\frac{Y_{10i}}{p_{10i}}-\frac{Y_{01i}}{p_{01i}}+\frac{Y_{00i}}{p_{00i}}\right)\frac{\partial p_{11i}}{\partial \theta}. \tag{2.2}$$

Here we have used that $\partial p_{10i}/\partial\theta = \partial(p_{1i}-p_{11i})/\partial\theta = -\partial p_{11i}/\partial\theta$, etc. Taking the expectation of the second-order derivatives yields

$$E\left\{\frac{\partial^2 l(\beta,\theta)}{\partial\beta^2}\right\} = -\sum\left\{\frac{1}{p_{11i}}\left(\frac{\partial p_{11i}}{\partial\beta}\right)\left(\frac{\partial p_{11i}}{\partial\beta}\right)' + \frac{1}{p_{10i}}\left(\frac{\partial p_{10i}}{\partial\beta}\right)\left(\frac{\partial p_{10i}}{\partial\beta}\right)' + \frac{1}{p_{01i}}\left(\frac{\partial p_{01i}}{\partial\beta}\right)\left(\frac{\partial p_{01i}}{\partial\beta}\right)' + \frac{1}{p_{00i}}\left(\frac{\partial p_{00i}}{\partial\beta}\right)\left(\frac{\partial p_{00i}}{\partial\beta}\right)'\right\},$$
$$E\left\{\frac{\partial^2 l(\beta,\theta)}{\partial\beta\,\partial\theta}\right\} = -\sum\left(\frac{1}{p_{11i}}\frac{\partial p_{11i}}{\partial\beta}-\frac{1}{p_{10i}}\frac{\partial p_{10i}}{\partial\beta}-\frac{1}{p_{01i}}\frac{\partial p_{01i}}{\partial\beta}+\frac{1}{p_{00i}}\frac{\partial p_{00i}}{\partial\beta}\right)\frac{\partial p_{11i}}{\partial\theta},$$
$$E\left\{\frac{\partial^2 l(\beta,\theta)}{\partial\theta^2}\right\} = -\sum\left(\frac{1}{p_{11i}}+\frac{1}{p_{10i}}+\frac{1}{p_{01i}}+\frac{1}{p_{00i}}\right)\left(\frac{\partial p_{11i}}{\partial\theta}\right)^2.$$

The dependence between a pair of observations can be quantified in different ways. Prentice (1988) considered the correlation between $Y_1$ and $Y_2$. This is intuitively very attractive but the correlation is restricted, and the restrictions depend on the marginals $p_1$ and $p_2$. Therefore, in the remainder of this paper we concentrate on two other measures, whose range does not depend on the marginals: the tetrachoric correlation and the odds ratio.

### 2.1. *Tetrachoric Correlation*

The use of the tetrachoric correlation as a measure of association in $2\times 2$ tables goes back to the beginning of this century (Pearson, 1900). Ashford and Sowden (1970) introduced the multivariate probit model, where the marginal probabilities may depend on certain covariates. We follow their approach but use logistic marginals instead of probit marginals.

The general idea is as follows. We suppose that the outcomes $(Y_1, Y_2)$ are realizations of a pair of latent continuous variables $(Z_1, Z_2)$, where $Z_1$ and $Z_2$ are bivariate standard normally distributed with correlation $\rho$. The outcome $Y_j$ $(j=1, 2)$ equals 1, if $Z_j < g_j$ with $g_j = \Phi^{-1}(p_j)$ where $\Phi$ is the standard normal cumulative distribution function. The marginal probabilities $p_j$ are logistic, $p_j = \exp(x_j\beta)/\{1+\exp(x_j\beta)\}$. This means that

$$p_{11} = \Pr(Z_1 < g_1, Z_2 < g_2) = \int_{-\infty}^{g_1}\int_{-\infty}^{g_2} f(t_1, t_2, \rho)\,dt_2\,dt_1.$$

In this expression, $f(t_1, t_2, \rho)$ is the joint density function of the standardized bivariate normal distribution, with correlation $\rho$. In this setting, the parameter $\rho$ is a measure

for the dependence between $Y_1$ and $Y_2$ and is called the tetrachoric correlation. In Stuart and Ord (1991) it is shown how $p_{11}$ can be evaluated by Hermite polynomials.

First-order derivatives of $p_{11}$ with respect to $\beta$ and $\rho$ are

$$\frac{\partial p_{11}}{\partial \beta} = \Phi\left\{\frac{g_2-\rho g_1}{\sqrt{(1-\rho^2)}}\right\}\frac{\partial p_1}{\partial \beta} + \Phi\left\{\frac{g_1-\rho g_2}{\sqrt{(1-\rho^2)}}\right\}\frac{\partial p_2}{\partial \beta}$$
$$\frac{\partial p_{11}}{\partial \rho} = f(g_1, g_2, \rho). \tag{2.3}$$

By combining this with the earlier general formulae, the derivation of a score statistic to test whether $\rho = 0$ is quite easy since

$$\left.\frac{\partial p_{11}}{\partial \rho}\right|_{\rho=0} = \phi(g_1)\,\phi(g_2),$$

with $\phi(g) = \{\sqrt{(2\pi)}\}^{-1}\exp(-1/2g^2)$, the univariate standard normal density function. This yields as score statistic $U$

$$U = \frac{\partial l(\beta, 0)}{\partial \rho} = \sum_{i=1}^{m} \frac{(Y_{1i}-p_{1i})(Y_{2i}-p_{2i})\,\phi(g_{1i})\,\phi(g_{2i})}{p_{1i}p_{2i}(1-p_{1i})(1-p_{2i})},$$

with asymptotic variance

$$\mathrm{var}(U) = E\left\{-\frac{\partial^2 l(\beta, 0)}{\partial \rho^2}\right\} = \sum_{i=1}^{m} \frac{\phi^2(g_{1i})\,\phi^2(g_{2i})}{p_{1i}p_{2i}(1-p_{1i})(1-p_{2i})},$$

and testing if $\rho = 0$ can be done by substituting the estimates obtained by ordinary logistic regression and comparing $U^2/\mathrm{var}(U)$ with a $\chi^2_{(1)}$-distribution.

The parameters can be obtained by the Newton–Raphson approach. We used the expected information matrix instead of the observed matrix of negative second derivatives. Our experience showed that in this way the iterations diverged less often. Furthermore, the expected information matrix is easier to compute. To remove the restriction $-1 < \rho < 1$, in our computations we used $\log\{(1+\rho)/(1-\rho)\}$. An estimate of the covariance matrix is obtained by taking the inverse of the expected information matrix and substituting the maximum likelihood estimates.

### 2.2. *Odds Ratio*

A second method to characterize the association among a $2 \times 2$ table is by means of the odds ratio. This measure is used by for example Dale (1986). She considers a more extended problem with correlated ordinal outcomes and defined the 'cross-ratio' to quantify the dependence between the outcomes. For binary responses the cross-ratio reduces to the odds ratio and with the notation of Table 1 the odds ratio $\psi$ is defined by

$$\psi = p_{11}p_{00}/p_{10}p_{01}.$$

An extensive description of the bivariate odds ratio model can be found in McCullagh and Nelder (1989). The odds ratio is easy to interpret: $\psi$ can be seen as the ratio of the odds of $Y_1 = 1$ given that $Y_2 = 1$ and the odds of $Y_1 = 1$ given that $Y_2 = 0$. If there is no dependence between $Y_1$ and $Y_2$, $\psi = 1$.

The joint probability $p_{11}$ can be expressed in terms of $p_1$, $p_2$ and $\psi$ as follows (Dale, 1986):

$$p_{11} = \begin{cases} 1/2(\psi-1)^{-1}\{1+(p_1+p_2)(\psi-1)-S(p_1,p_2,\psi)\} & \text{if } \psi \neq 1, \\ p_1 p_2 \qquad \text{if } \psi = 1 \end{cases} \tag{2.4}$$

where

$$S(p_1,p_2,\psi) = \sqrt{[\{1+(p_1+p_2)(\psi-1)\}^2 + 4\psi(1-\psi)p_1 p_2]}.$$

The odds ratio satisfies $\psi \geqslant 0$ and to remove this restriction in the computations $\Delta = \log \psi$ is used. The first derivative of $p_{11}$ with respect to $\beta$ and $\Delta$ are (Palmgren, 1989; McCullagh and Nelder, 1989)

$$\frac{\partial p_{11}}{\partial \beta} = \left(\frac{1}{p_{11}} + \frac{1}{p_{10}} + \frac{1}{p_{01}} + \frac{1}{p_{00}}\right)^{-1}\left\{\frac{\partial p_1}{\partial \beta}\left(\frac{1}{p_{10}} + \frac{1}{p_{00}}\right) + \frac{\partial p_2}{\partial \beta}\left(\frac{1}{p_{01}} + \frac{1}{p_{00}}\right)\right\},$$

$$\frac{\partial p_{11}}{\partial \Delta} = \left(\frac{1}{p_{11}} + \frac{1}{p_{10}} + \frac{1}{p_{01}} + \frac{1}{p_{00}}\right)^{-1}.$$

Combining these expressions with expressions (2.2) it follows that the first derivative of the log-likelihood with respect to $\beta$ depends only on the marginal outcomes $Y_1$ and $Y_2$ but not on the joint outcome $Y_1 Y_2$. Nevertheless, the normal equations for $\beta$ differ from the situation where $Y_1$ and $Y_2$ are independent.

Estimation of the parameters can be done by Newton–Raphson iteration. A score test for testing whether the odds ratio is 1 or equivalently $\Delta = 0$ is also easy to derive:

$$W = \frac{\{\partial l(\beta,0)/\partial \Delta\}^2}{E\{-\partial^2 l(\beta,0)/\partial \Delta^2\}} = \frac{\left\{\sum_{i=1}^{m}(Y_{1i}-p_{1i})(Y_{2i}-p_{2i})\right\}^2}{\sum_{i=1}^{m} p_{1i}p_{2i}(1-p_{1i})(1-p_{2i})}.$$

This test statistic is intuitively more appealing than the test statistic for $\rho = 0$ because it is directly based on the sum of the cross-products per cell $(Y_1 - p_1)(Y_2 - p_2)$. It is clear that if there is no dependence the expectation of each cross-product equals 0. This is revealed in the test statistic: the sum of the cross-products also has expected value 0 if $\Delta = 0$ and the test statistic $W$, the squared sum of the cross-products over the variance of the sum of the cross-products, is then asymptotically $\chi^2_{(1)}$ distributed. Hence, if there is no dependence, we would expect $W$ to be around 1, whereas if there is dependence we expect $W$ to be larger.

### 2.3. *Comparison of Tetrachoric Correlation and Odds Ratio*

One of the major advantages of the odds ratio is its direct interpretation. Furthermore it is quite easy to allow the odds ratio to depend on certain covariates by assuming that $\log \psi = x\gamma$ with $x$ a vector of covariates and $\gamma$ a parameter vector. The tetrachoric correlation can depend on $x$ by $\rho = x\gamma$ but here problems arise because $\rho$ is restricted to between $+1$ and $-1$. Another approach is to let $\log\{(1+\rho)/(1-\rho)\}$ depend linearly on $x$ but this yields parameters which are difficult to interpret.

The tetrachoric correlation model has the advantage that it can be naturally

extended to an arbitrary number of correlated observations (see Section 4). The odds ratio model can be extended to a general multivariate logistic model as is described in McCullagh and Nelder (1989), by considering the logistic transforms of the marginal probabilities combined with multivariate logistic contrasts (see also Liang *et al.* (1992)). But for this model it is generally not possible to derive closed expressions for the joint probabilities like equations (2.4) which makes it difficult to compute the full likelihood.

We can compare the two methods by considering the first-order approximations of $\hat{\rho}$ and $\hat{\Delta}$ in the neighbourhood of 0. This yields that $\hat{\Delta}$ is approximately

$$\hat{\Delta} = \frac{\partial l(\beta, 0)/\partial \Delta}{E\{-\partial^2 l(\beta, 0)/\partial \Delta^2\}} = \frac{\sum_{i=1}^{m} (Y_{1i} - p_{1i})(Y_{2i} - p_{2i})}{\sum_{i=1}^{m} p_{1i} p_{2i}(1 - p_{1i})(1 - p_{2i})},$$

and that a first-order estimate of $\hat{\rho}$ is given by

$$\hat{\rho} = \frac{\partial l(\beta, 0)/\partial \rho}{E\{-\partial^2 l(\beta, 0)/\partial \rho^2\}} = \frac{\sum_{i=1}^{m} (Y_{1i} - p_{1i})(Y_{2i} - p_{2i})\, w(p_{1i})\, w(p_{2i})}{\sum_{i=1}^{m} p_{1i} p_{2i}(1 - p_{1i})(1 - p_{2i})\, w^2(p_{1i})\, w^2(p_{2i})},$$

with

$$w(p_j) = \phi\{\Phi^{-1}(p_j)\}/p_j(1 - p_j).$$

Both approximations are weighted averages of the cross-products per cell $(Y_1 - p_1)(Y_2 - p_2)$. The expression for $\hat{\rho}$ is a weighted version of the expression for $\hat{\Delta}$ with weights $w(p_{1i})\, w(p_{2i})$. Fig. 1 is a plot of $w(p)$ as function of $p$. For $p$-values between 0.2 and 0.8, $w(p)$ is virtually constant at about 1.7. This is the same heuristic factor which appears when we compare the probit model with the logistic model: the coefficients of the logistic model are about 1.7 times larger than the coefficients of the probit model. If the marginal probabilities are not too close to 1 or 0, the relationship between $\hat{\Delta}$ and $\hat{\rho}$ can be given by

$$\hat{\Delta} = (1.7)^2 \hat{\rho}, \tag{2.5}$$

and both approaches will yield more or less the same results.

## 3. Example (Part I)

As an example we considered the 107 pairs of complete twins of the POPS data set. As outcome variable $Y$ we used neonatal mortality, where $Y = 1$ if an infant dies within 28 days. Several covariates were considered, such as gestational age, birthweight, and some 1–0 variables like sex (male/female), prolonged rupture of membranes (yes/no), use of corticosteroids (yes/no), fetal position (breech/vertex), the mode of delivery (normal/section) and the pre-existence of maternal diseases (yes/no). The crude estimates of the dependence between twins were $\hat{\rho} = 0.702$

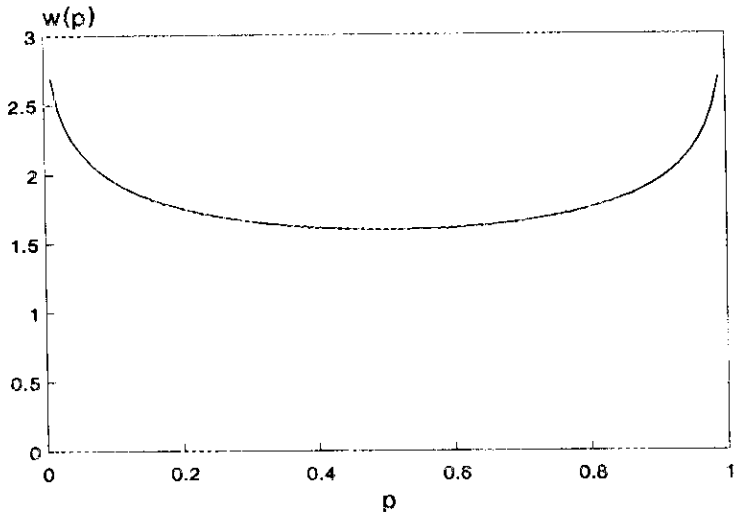


Fig. 1. Plot of weight $w(p) = \phi\{\Phi^{-1}(p)\}/p(1-p)$ against $p$

(standard error 0.100), $\log \psi = 1.999$ (standard error 0.210) and $\psi = 7.38$, and we see, not surprisingly, that the outcomes between twins are highly correlated.

What is of interest is whether the apparent dependence can be explained by some of the covariates. Therefore we start by modelling the marginal probabilities with a rather simple logistic model with linear terms in gestational age and birthweight. Since both variables are strong predictors for survival and are very similar within a pair of twins, we expect that the dependence would reduce. Table 2 shows that this is the case. The estimates of $\beta$ in the odds ratio model and the tetrachoric correlation model differ only slightly. This phenomenon that the marginal parameters are not very sensitive to the choice of dependence parameter has also been found in simulation studies by Lipsitz *et al.* (1990). The log-likelihood of the tetrachoric model is somewhat larger than the log-likelihood of the odds ratio model but the difference is

TABLE 2
*Parameter estimates and standard errors for the tetrachoric correlation model, the bivariate odds ratio model and under the assumption of independence*†

| *Parameter* | *Without correlation* | | *Tetrachoric correlation* | | *Odds ratio* | |
|---|---|---|---|---|---|---|
| | *Estimate* | *Standard error* | *Estimate* | *Standard error* | *Estimate* | *Standard error* |
| Intercept | 17.70 | 3.25 | 17.69 | 3.50 | 17.72 | 3.50 |
| Gestational age | −0.600 | 0.130 | −0.596 | 0.138 | −0.599 | 0.138 |
| Birthweight | −0.0805 | 0.0824 | −0.0891 | 0.0844 | −0.0856 | 0.0842 |
| Tetrachoric correlation | — | — | 0.377 | 0.171 | — | — |
| Δ (log-odds ratio) | — | — | — | — | 1.075 | 0.535 |
| Log-likelihood | −103.98 | | −101.83 | | −101.97 | |

†The marginals are logistic with birthweight (in 100 g) and gestational age (in weeks) linear.

not remarkable. Clearly the standard errors obtained under the assumption of independence are smaller. This is what we would expect since gestational age and birthweight vary between the different twin pairs but within a pair of twins the gestational ages are the same and the birthweights are very similar. Ignoring the dependence, one assumes more information in the data about gestational age and birthweight than there really is. This usually yields too small estimates of the standard errors. To illustrate this, in the extreme situation where there is perfect correlation, i.e. within a twin the two infants are exactly identical, acting as if the outcomes within a twin are independent yields standard errors which are $\sqrt{2}$ times too small. Approximation (2.5) satisfies well: $(1.7)^2\hat{\rho} = 1.088$, which is close to the value of $\log \hat{\psi} = 1.075$ of Table 2.

First fitting the marginal probabilities with standard logistic regression and then maximizing the bivariate log-likelihood function (2.1) with the marginals fixed is easier to perform. This approach yielded virtually the same results for both models (tetrachoric correlation model, log-likelihood $-101.84$; odds ratio model, log-likelihood $-101.98$).

Table 2 shows that an infant's odds of dying within 28 days, corrected for birthweight and gestational age, will be 2.9 times as large if its twin brother or sister dies than if its sibling stays alive.

The score statistic for $\rho = 0$ was equal to 4.79 ($p = 0.03$) and score test for $\psi = 1$ equalled 4.24 ($p = 0.04$), so there is still a significant dependence between twins. We considered the effect of the other covariates and also examined quadratic terms in gestational age and birthweight and some interaction terms. None contributed significantly to the model. For example, logistic marginals with all explanatory variables linear in the model yielded virtually the same value of the log-likelihood, whereas the estimates for the dependence were even slightly larger (odds ratio model, $\hat{\Delta} = 1.09$ (standard error 0.54), log-likelihood $-101.39$; tetrachoric correlation model, $\hat{\rho} = 0.38$ (standard error 0.17), log-likelihood $-101.47$).

The dependence could be higher for identical twins. Unfortunately the zygoticity of a twin was not recorded, but we looked at whether the dependence was higher for twins of the same sex. Therefore the dependence parameters were allowed to depend on an indicator variable $z$ which takes value 1 if both twins are of the same sex and 0 otherwise. Fitting the tetrachoric correlation model with $\log\{(1+\rho)/(1-\rho)\} = \gamma_0 + \gamma_1 z$ yields $\hat{\gamma}_0 = 0.715$ (standard error 0.806), $\hat{\gamma}_1 = 0.102$ (standard error 0.927) and log-likelihood $-101.82$. This gives a tetrachoric correlation of 0.343 for twins of different sex, whereas for twins of the same sex $\hat{\rho} = 0.387$. Although the dependence is somewhat larger for like-sexed twins, the differences are not significant at all. The same was true for the odds ratio model. Defining $\Delta = \log \psi = \gamma_0 + \gamma_1 z$ yields $\hat{\gamma}_0 = 1.043$ (standard error 1.097), $\hat{\gamma}_1 = 0.044$ (standard error 1.257) and log-likelihood $-101.97$.

The dependence between observations changes the estimates of the joint probabilities. To visualize this, contour plots of $p_{11}$ as a function of $p_1$ and $p_2$ are made for the three models of Table 2 and plotted in Fig. 2, which shows that the estimates of $p_{11}$ obtained by the two dependence models differ only slightly, whereas the values obtained under the assumption of independence are considerably smaller. For example, if $(p_1, p_2) = (0.5, 0.5)$, the joint probability under the assumption of independence $p_{11}$ will be exactly 0.25, whereas $p_{11}$ is larger for the two dependence models: the point (0.5, 0.5) is between the contour lines $p_{11} = 0.25$ and $p_{11} = 0.5$. (To be

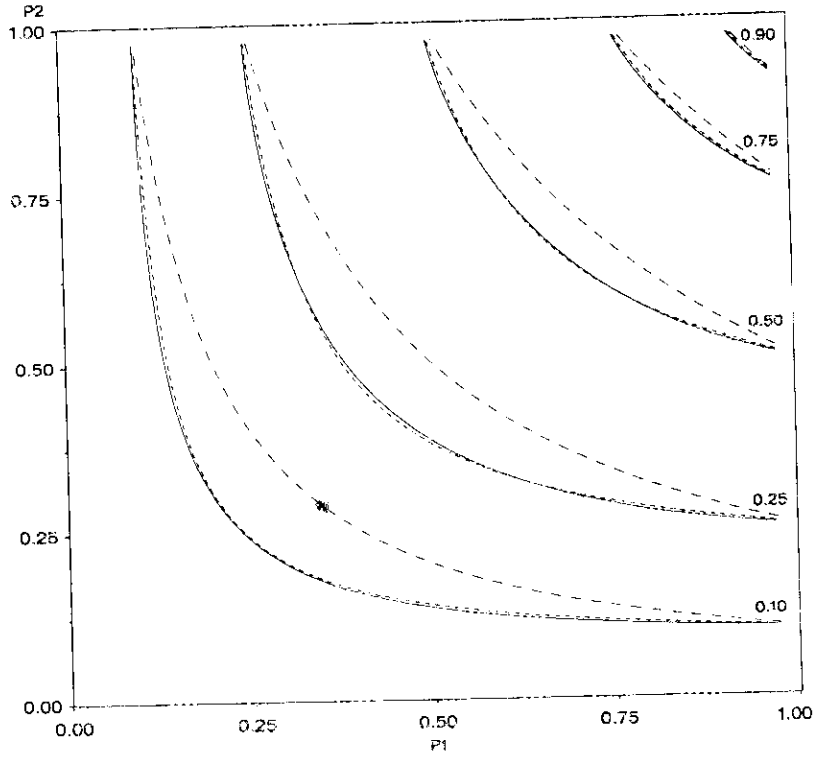


Fig. 2. Contour plots of the joint probability $p_{11}$ as a function of $p_1$ and $p_2$ for the tetrachoric model (———), the odds ratio model (--------) and under the assumption of independence (------): the tetrachoric correlation equals 0.377, the log-odds ratio equals 1.075, values obtained for the data set of complete pairs of twins (see Table 2); contour lines, i.e. lines with constant value of $p_{11}$, correspond to the $p_{11}$-values 0.10, 0.25, 0.5, 0.75 and 0.9, as indicated

precise, the tetrachoric model yielded a fitted value of $\hat{p}_{11} = 0.312$; the odds ratio model yielded $\hat{p}_{11} = 0.316$.)

## 4. Extension to Arbitrary Number of Correlated Observations per Block

In this section the tetrachoric approach is extended to an arbitrary number of dependent observations. Consider data from $m$ blocks, and let $k_i$ be the number of observations in block $i$. We suppose that the observations in a block are correlated and observations in different blocks are independent. As before, an underlying structure of continuous variables $Z_1, \ldots, Z_{k_i}$ is assumed, such that for a block of size $k_i$ with outcomes $Y_1, \ldots, Y_{k_i}$ the outcome $Y_j = 1$ $(j = 1, \ldots, k_i)$ if $Z_j < \Phi^{-1}(p_j)$, with $p_j = \Pr(Y_j = 1)$ the marginal probability of $Y_j$ being 1. The marginals are logistic, $p_j = \exp(x_j\beta)/\{1 + \exp(x_j\beta)\}$, and the latent variables $Z_1, \ldots, Z_{k_i}$ are multivariate normally distributed with mean 0, variance 1 and all correlations equal to $\rho$. The case of different correlation coefficients is not considered here but can be treated in the same way.

The contribution of a block to the true log-likelihood function is

$$\sum_{j_1=0}^{1} \cdots \sum_{j_{k_i}=0}^{1} Y_{j_1 \ldots j_{k_i}} \log p_{j_1 \ldots j_{k_i}}.$$

In this expression $Y_{j_1 \ldots j_{k_i}} = 1$ if $Y_1 = j_1 \ldots Y_{k_i} = j_{k_i}$ and $Y_{j_1 \ldots j_{k_i}} = 0$ otherwise, and $p_{j_1 \ldots j_{k_i}} = \Pr(Y_1 = j_1 \ldots Y_{k_i} = j_{k_i})$. The probabilities that all $Y_j = 1$ is given by

$$p_{1 \ldots 1} = \int_{-\infty}^{g_1} \cdots \int_{-\infty}^{g_{k_i}} f(t_1, \ldots, t_{k_i}, \rho)\, dt_{k_i} \ldots dt_1,$$

with $f(t_1, \ldots, t_k, \rho)$, the $k$-variate standard normal density function with all correlations equal to $\rho$ and with $g_j = \Phi^{-1}(p_j)$.

To evaluate the likelihood of these data, for all blocks, $p_{j_1 \ldots j_{k_i}}$ has to be computed for all combinations of $j_1 \ldots j_{k_i}$ which is very difficult to perform when the block sizes increase. Ochi and Prentice (1984) consider an equicorrelated probit model with equal success probabilities in each block and obtain approximations for the joint probabilities. But even then, the computation can be quite burdensome.

In recent years there has been an increased use of estimation techniques which do not require knowledge about the whole underlying distribution of the data. Basically the idea is to approximate the true likelihood by a function that is easier to evaluate, which is termed pseudolikelihood (Besag, 1975) or quasi-likelihood (McCullagh, 1983), depending on the type of function chosen, or to replace the true normal equations by equations that are easier to solve, the so-called estimation equations. Applications of pseudolikelihood estimation can be found in Azzalini (1983) and in Connolly and Liang (1988). Generalized estimation equations are used by Zeger and Liang (1986) and by Prentice (1988). For a discussion of these methods see Liang *et al.* (1992). Usually these methods still yield consistent estimators as long as the first derivatives of the pseudolikelihood (the estimation functions) have mean 0 at the true parameter values. The estimators will not always be efficient, however.

We focus here on the likelihood function itself, which we replace by a function that is easy to evaluate which takes its maximum at the same place. This is done by using the product of all pairwise likelihoods in a block instead of the true contribution of a block to the likelihood. In our POPS example that would mean that we treat a triplet as if it consists of three pairs of twins etc. The pseudolikelihood constructed in this way requires only the evaluation of bivariate normal integrals and is therefore much easier to evaluate than the full likelihood function. The contribution of block $i$ to the pseudo-log-likelihood is

$$l_i = Y \log p + (1 - Y) \log(1 - p),$$

if the block contains only one observation. Otherwise, if the block size $k_i > 1$, the contribution is

$$l_i = \frac{1}{k_i - 1} \sum_{p=1}^{k_i} \sum_{q=1}^{p-1} l_{pq},$$

with $l_{pq}$ the pairwise likelihood of $(Y_p, Y_q)$, with $Y_p$ and $Y_q$ responses in block $i$:

$$l_{pq} = \sum_{j_p=0}^{1} \sum_{j_q=0}^{1} Y_{j_p j_q} \log p_{j_p j_q}.$$

Here $Y_{j_p j_q} = 1$ if $Y_p = j_p$ and $Y_q = j_q$, and $Y_{j_p j_q} = 0$ otherwise, and $p_{j_p j_q} = \Pr(Y_p = j_p, Y_q = j_q)$.

If all blocks have size 1 or 2, pseudolikelihood and maximum likelihood coincide. In each block the summation is done over $k_i(k_i - 1)/2$ terms $l_{pq}$, each term corresponding to two observations. By dividing the sum of the $l_{pq}$ by $k_i - 1$ the contribution of each block to the likelihood is relative to its block size. In this way the pseudolikelihood equals the full likelihood if $\rho = 0$; hence for $\rho = 0$ the parameters are estimated efficiently. Therefore we do not expect a great loss in efficiency by using the pseudolikelihood if $\rho$ is not too far from 0.

The total pseudo-log-likelihood is the sum of the $l_i$ and the pseudolikelihood estimators $(\tilde{\beta}, \tilde{\rho})$ are obtained by solving

$$U(\beta, \rho) = \sum_{i=1}^{m} u_i(\beta, \rho) = \sum_{i=1}^{m} \frac{\partial l_i(\beta, \rho)}{\partial(\beta, \rho)} = 0. \tag{4.1}$$

If the mean and the covariance matrix of the response vector are specified correctly (i.e. if $E(Y_{j_p j_q}) = p_{j_p j_q}$, for all $j_p$ and $j_q$), it follows directly that $E\{U(\beta_0, \rho_0)\} = 0$, with $(\beta_0, \rho_0)$ the true parameter values. Then it is straightforward to show that the pseudolikelihood estimators are consistent, by using the same arguments as in the proof of the consistency of maximum likelihood estimators (see for example Stuart and Ord (1991)). Under some regularity conditions, the distribution of $(\tilde{\beta} - \beta_0, \tilde{\rho} - \rho_0)\sqrt{m}$ converges to an $N(0, \Lambda)$ distribution with

$$\Lambda = mH_1^{-1}H_2H_1^{-1},$$

where

$$H_1 = I(\beta_0, \rho_0) = E\{\partial U(\beta_0, \rho_0)/\partial(\beta, \rho)\},$$

$$H_2 = \operatorname{cov}\{U(\beta_0, \rho_0)\},$$

and an estimator $\hat{\Lambda}$ is obtained by substituting

$$\hat{H}_1 = I(\tilde{\beta}, \tilde{\rho}),$$

$$\hat{H}_2 = \sum_{i=1}^{m} u_i(\tilde{\beta}, \tilde{\rho})\, u_i(\tilde{\beta}, \tilde{\rho})'.$$

This so-called 'sandwich estimator' $\hat{\Lambda}$ is the robust variance estimate of Zeger and Liang (1986). Royall (1986) discussed general properties of this type of estimator and gave several applications. Note that the dependence structure does not need to be expanded to compute $\hat{H}_2$.

In the proof of the consistency of $(\tilde{\beta}, \tilde{\rho})$, only assumptions about the pairwise dependence between the observations are made but no assumptions about the multivariate dependence structure. Therefore, in the same way, the odds ratio can also be used to characterize the pairwise dependence in a multivariate binary regression model. The advantage of the tetrachoric correlation is that, after obtaining the pseudolikelihood estimates of $\beta$ and $\rho$, estimates of the multivariate joint probabilities can be obtained by a single evaluation of a multinormal integral.

## 5. Example (Part II)

We consider now the total POPS data set, consisting of 1338 infants. After deleting the observations with missing data, a data set of 1335 infants remains, consisting of 1024 singletons and 311 infants of multiples. There were 276 twins: 214 infants from complete pairs of twins and 62 infants without a twin brother or sister in the study. Of the 34 infants from triplets, 15 came from complete triplets, 12 had one triplet brother or sister in the study and 7 were single. There was one quadruplet infant in the study. These data were analysed by using the pseudolikelihood approach of Section 4 and compared with the results of a standard logistic regression. The tetrachoric correlation was assumed to be equal within all pairs of twins and triplets. The same covariates were considered as in Section 3, together with an extra variable which indicates whether an infant results from multiple birth. To prevent numerical problems the (rounded) means were subtracted from the covariates in the quadratic terms. A forward selection procedure was used and the estimates corresponding to the significant variables are given in Table 3.

More terms were significant than in the analysis of the complete pairs of twins. This can be explained by the larger sample size and by the larger ranges of the continuous covariates. Several singletons have a very high or very low birthweight, relative to their gestational age. These infants are either growth retarded or extremely heavy for their age and these infants have a lower survival probability. Quadratic terms in gestational age and birthweight are needed to obtain adequate predictions for these infants. In the subset of complete twins there are no infants with such extreme birthweights and there a model with only linear terms is sufficient. More details about modelling the continuous covariates of these data and the fit of various models can be found in le Cessie and van Houwelingen (1991).

The estimates of $\beta$ in the two models differ slightly. The estimated correlation coefficient was 0.458, even somewhat larger than in the analysis of the subset of complete pairs of twins, and a Wald-type test ($\hat{\rho}/\widehat{\text{se}}(\hat{\rho})$) yields a $p$-value of 0.001.

To our surprise the estimated standard errors in the tetrachoric correlation model were somewhat smaller for almost all variables. As in the subset of complete twins we

TABLE 3

*Parameter estimates and standard errors for the tetrachoric correlation model and under the assumption of independence*†

| *Parameter* | *Without correlation* | | *Tetrachoric correlation* | |
|---|---|---|---|---|
| | *Estimate* | *Standard error* | *Estimate* | *Standard error* |
| Intercept | 9.23 | 1.02 | 9.14 | 1.00 |
| Gestational age (weeks) | −0.376 | 0.036 | −0.373 | 0.036 |
| Birthweight (100 g) | −0.0017 | 0.0314 | −0.0015 | 0.0292 |
| (Gestational age − 30)$^2$ | 0.0322 | 0.0087 | 0.0320 | 0.0074 |
| (Birthweight − 15)$^2$ | 0.0202 | 0.0045 | 0.0202 | 0.0041 |
| Corticosteroids | −0.679 | 0.234 | −0.717 | 0.230 |
| Multiple pregnancy | 0.461 | 0.176 | 0.431 | 0.186 |
| Fetal position | 0.341 | 0.166 | 0.341 | 0.156 |
| Tetrachoric correlation | — | — | 0.458 | 0.153 |

†The marginals are logistic.

would expect an increase in the standard errors since most of the covariates are the same or very similar within a multiple. The smaller estimated standard errors could be caused by the fact that two different estimators of the standard errors are considered, and in each estimate there is a certain variability. Thus it could be because of random variation that we observe smaller standard errors. Furthermore, although the estimators of the standard errors are consistent under general conditions (Royall, 1986), in finite samples the estimator $\hat{\Lambda}$ tends to underestimate the covariance matrix, because estimates are substituted in the matrix $\hat{H}_2$. This is similar to what also appears when estimating the variance of residuals $Y_i - p_i$. If $p_i$ is replaced by its maximum likelihood estimator $\hat{p}_i$, the variance of $Y_i - \hat{p}_i$ is approximately $(1 - h_i)p_i(1 - p_i)$, with $h_i$ the diagonal element of the hat matrix. In our example, the shrinkage effect is more complicated and it is not easy to give an explicit estimate of its magnitude.

## 6. Discussion

Pseudolikelihood estimation and related estimation techniques are very helpful if the full underlying distribution of the data is unknown or if the true likelihood is difficult to evaluate, and nowadays in an increasing number of applications these methods are used. Our method resembles the approach of Zeger and Liang (1986) and Prentice (1988), but the pseudolikelihood method is somewhat easier to understand and the multivariate tetrachoric correlation model has the advantage that estimates of the joint probabilities can be generated relatively easily. More research is needed to determine the loss in efficiency by using a pseudolikelihood. Since the pseudolikelihood equals the full likelihood for $\rho = 0$, we expect only small losses when $\rho$ is small.

In our example, the estimates of $\beta$ and its standard error did not change considerably in the dependence models. But dependence models are needed to study the magnitude of the dependence and its relation with the covariates. Furthermore these models yield correct estimates of the joint probabilities.

We derived score tests for independence for two correlated observations. Deriving the exact score test for more dimensions is difficult but by using the pseudolikelihood a pseudoscore test can be derived straightforwardly. The tests derived in this way have forms that are similar to the score tests of Section 2, with summation over all possible combinations $(Y_{pi} - p_{pi})(Y_{qi} - p_{qi})$ in each block.

We focused on binary outcomes in the POPS data. For most infants the exact survival times are known and bivariate and multivariate survival models could also be applied. One way of modelling multivariate survival data is by means of frailty models (Clayton and Cuzick, 1985; Hougaard, 1986; Nielsen *et al.*, 1992). These models can be seen as generalizations of the random effect models for binary data, and the parameter estimates do not have a marginal interpretation. Models based on the marginal survival probabilities are more in the line of our paper.

**Acknowledgement**

We would like to thank the referees for their helpful comments on earlier versions of this paper.

## References

Ashford, J. R. and Sowden, R. R. (1970) Multivariate probit analysis. *Biometrics*, **26**, 535–546.

Azzalini, A. (1983) Maximum likelihood estimation of order $m$ for stationary stochastic processes. *Biometrika*, **70**, 381–387.

Besag, J. E. (1975) The statistical analysis of non-lattice data. *Statistician*, **24**, 179–195.

le Cessie, S. and van Houwelingen, J. C. (1991) A goodness of fit test for binary regression models based on smoothing methods. *Biometrics*, **47**, 1267–1282.

Clayton, D. and Cuzick, J. (1985) Multivariate generalizations of the proportional hazards model (with discussion). *J. R. Statist. Soc.* A, **148**, 82–117.

Connolly, M. A. and Liang, K.-Y. (1988) Conditional logistic regression models for correlated binary data. *Biometrika*, **75**, 501–506.

Dale, J. R. (1986) Global cross-ratio models for bivariate, discrete, ordered responses. *Biometrics*, **42**, 909–917.

Hougaard, P. (1986) Survival models for heterogeneous populations derived from stable distributions. *Biometrika*, **73**, 387–396.

Liang, K.-Y., Zeger, S. L. and Qaqish, B. (1992) Multivariate regression analyses for categorical data (with discussion). *J. R. Statist. Soc.* B, **54**, 3–40.

Lipsitz, S. R., Laird, N. M. and Harrington, D. P. (1990) Maximum likelihood regression methods for paired binary data. *Statist. Med.*, **9**, 1517–1525.

McCullagh, P. (1983) Quasi likelihood functions. *Ann. Statist.*, **11**, 59–67.

McCullagh, P. and Nelder, J. A. (1989) *Generalized Linear Models*, 2nd edn. London: Chapman and Hall.

Neuhaus, J. M., Kalbfleisch, J. D. and Hauck, W. W. (1991) A comparison of cluster-specific and population-averaged approaches for analyzing correlated binary data. *Statist. Rev.*, **59**, 25–36.

Nielsen, G. G., Andersen, P. K., Gill, R. D. and Sorensen, T. I. A. (1992) A counting process approach in maximum likelihood estimation in frailty models. *Scand. J. Statist.*, **19**, 25–43.

Ochi, Y. and Prentice, R. L. (1984) Likelihood inference in a correlated probit regression model. *Biometrika*, **71**, 531–543.

Palmgren, J. (1989) Regression models for bivariate binary responses. *Technical Report 101*. Department of Biostatistics, School of Public Health and Community Medicine, Seattle.

Pearson, K. (1900) Mathematical contribution to the theory of evolution: VII, On the correlation of characters not quantitatively measurable. *Phil. Trans. R. Soc. Lond.* A, **195**, 1–47.

Prentice, R. L. (1988) Correlated binary regression with covariates specific to each binary observation. *Biometrics*, **44**, 1033–1048.

Royall, R. M. (1986) Model robust confidence intervals using maximum likelihood estimators. *Int. Statist. Rev.*, **54**, 221–226.

Stuart, A. and Ord, J. K. (1991) *Kendall's Advanced Theory of Statistics*, vol. 2, p. 659. London: Arnold.

Verloove, S. P. and Verwey, R. Y. (1988) Project on preterm and small-for-gestational age infants in the Netherlands, 1989. *Thesis*. University of Leiden, Leiden. (Available from University Microfilms International, Ann Arbor, MI, USA, no. 8807276.)

Zeger, S. L. and Liang, K.-Y. (1986) Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, **42**, 1019–1031.

Zeger, S. L., Liang, K.-Y. and Albert, P. S. (1988) Models for longitudinal data: a generalized estimation equation approach. *Biometrics*, **44**, 1049–1060.