Biometrika (1997), **84**, 3, *pp.* 609–618
Printed in Great Britain

A goodness-of-fit test for logistic regression models based on case-control data

BY JING QIN

Department of Mathematics, University of Maryland, College Park, Maryland 20742, U.S.A.
e-mail: jqin@math.umd.edu

AND BIAO ZHANG

Department of Mathematics, University of Toledo, Toledo, Ohio 43606, U.S.A.
e-mail: bzhang@math.utoledo.edu

SUMMARY

We test the logistic regression assumption under a case-control sampling plan. After reparameterisation, the assumed logistic regression model is equivalent to a two-sample semiparametric model in which the log ratio of two density functions is linear in data. By identifying this model with a biased sampling model, we propose a Kolmogorov–Smirnov-type statistic to test the validity of the logistic link function. Moreover, we point out that this test statistic can also be used in mixture sampling. We present a bootstrap procedure along with some results on simulation and on analysis of two real datasets.

Some key words: Biased sampling problem; Bootstrap resampling; Case-control data; Goodness-of-fit test; Kolmogorov–Smirnov two-sample statistic; Logistic regression; Prospective sampling; Retrospective sampling.

1. INTRODUCTION

Logistic regression models are commonly used in analysing binary data which arise in studying relationships between diseases and environment or genetic characteristics; see for example Breslow & Day (1980, Ch. 4), Prentice & Pyke (1979) and Farewell (1979). In the set-up of a generalised linear model, Pregibon (1980) considered examining the adequacy of the hypothesised link for a given prospective sampling dataset. A non-parametric regression method was proposed by Azzalini, Bowman & Hardle (1989) to test the validity of the logistic regression assumption. Some graphical methods for assessing logistic regression models were studied by Landwehr, Pregibon & Shoemaker (1984). Let y be a binary response variable and x be the associated $1 \times p$ covariate. Then the standard logistic regression model has the form

$$P(y=1|x)=\frac{\exp(\alpha^*+x\beta)}{1+\exp(\alpha^*+x\beta)}\equiv\psi(x), \tag{1·1}$$

where α^* is a scale parameter and β is a $p\times 1$ vector parameter. In model (1·1), the marginal distribution of x is unspecified. In this paper, we consider testing the assumption of model (1·1) on the basis of case-control data. By case-control data, also called retrospective sampling data (Prentice & Pyke, 1979), we mean two independent groups of sample data as specified below.

Let $x_1,\ldots,x_{n_0}$ be a random sample from $F(x|y=0)$ and, independent of the x_i, let

$z_1, \ldots, z_{n_1}$ be a random sample from $F(x \mid y=1)$. Let $n = n_0 + n_1$ and let

$$\{t_1, t_2, \ldots, t_n\} = \{x_1, \ldots, x_{n_0}; z_1, \ldots, z_{n_1}\}$$

denote the combined sample. We assume that $n_i/n \to \rho_i > 0$ as $n \to \infty$ for $i = 0, 1$. If

$$\pi = P(y=1) = 1 - P(y=0)$$

and $f(x \mid y=i) = dF(x \mid y=i)/dx$ represents the conditional density or frequency function of x given $y=i$ for $i=0, 1$, then Bayes' rule gives

$$f(x \mid y=1) = \frac{\psi(x)}{\pi} f(x), \quad f(x \mid y=0) = \frac{1-\psi(x)}{1-\pi} f(x),$$

where $f(x)$ is the marginal distribution of x. It is seen that

$$\frac{f(x \mid y=1)}{f(x \mid y=0)} = \frac{1-\pi}{\pi} \frac{\psi(x)}{1-\psi(x)}.$$

Let $g(x) = f(x \mid y=0)$ and $h(x) = f(x \mid y=1)$ and let $G(x)$ and $H(x)$ be their corresponding cumulative distribution functions. Then we have

$$h(x) = f(x \mid y=1) = \frac{1-\pi}{\pi} \frac{\psi(x)}{1-\psi(x)} g(x) = \exp(\alpha + x\beta) g(x),$$

where $\alpha = \alpha^* + \log\{(1-\pi)/\pi\}$. As a result, we arrive at the following two-sample semiparametric model in which $(x_1, \ldots, x_{n_0})$ and $(z_1, \ldots, z_{n_1})$ are independent and

$$\begin{aligned} &x_1, \ldots, x_{n_0} \text{ are independent with density } g(x), \\ &z_1, \ldots, z_{n_1} \text{ are independent with density } h(x) = \exp(\alpha + x\beta) g(x). \end{aligned} \tag{1·2}$$

Note that model (1·2) is a biased sampling model with weight function $\exp(\alpha + x\beta)$ depending on the unknown parameters α and β. Vardi (1982, 1985) and Gill, Vardi & Wellner (1988) have discussed the biased sampling problem in the case of completely known weight function.

Our focus of attention in this paper is to test the validity of model (1·2). In the special case of $\alpha = \beta = 0$, we are led to test the equality of G and H for which a commonly-used test statistic is the Kolmogorov–Smirnov two-sample statistic KS given by

$$\text{KS} = \sup_t |\hat{G}(t) - \hat{H}(t)| = \frac{n}{n_1} \sup_t |\hat{G}(t) - \tilde{G}_0(t)|, \tag{1·3}$$

where

$$\hat{G}(t) = \frac{1}{n_0} \sum_{i=1}^{n_0} I(x_i \leqslant t), \quad \hat{H}(t) = \frac{1}{n_1} \sum_{j=1}^{n_1} I(z_j \leqslant t), \quad \tilde{G}_0(t) = \frac{1}{n} \sum_{i=1}^{n} I(t_i \leqslant t).$$

Note that $\hat{G}$ and $\hat{H}$ are, respectively, the nonparametric maximum likelihood estimators of G and H without the assumption of $G(t) = H(t)$, whereas $\tilde{G}_0$ is the nonparametric maximum likelihood estimator of G with the assumption of $G(t) = H(t)$. In the general case of $\beta \neq 0$, (1·3) motivates us to employ the statistic

$$\Delta = n^{\frac{1}{2}} \sup_t |\hat{G}(t) - \tilde{G}(t)|$$

to test the validity of model (1·2), where $\tilde{G}$ is the maximum semiparametric likelihood estimator of G under model (1·2). The explicit definition of $\tilde{G}$ is given in § 2.

This paper is organised as follows, In § 2, we propose our test statistic and present the main results. In § 3, we propose a bootstrap method for finding P-values of the proposed test. Also in § 3, we report results of simulation studies and two real data problems. Some discussion is given in § 4. The main theoretical results are proved in the Appendix.

2. Test statistics and main results

As mentioned earlier, there are two ways of estimating $G(t)$. One is to use the empirical distribution function based on the first sample, and the other is to use both samples by exploiting (1·2). The discrepancy between the two estimators will then allow us to assess the validity of model (1·2). In this section, we estimate $G(t)$ under model (1·2) by maximising the semiparametric likelihood $\mathscr{L}(\alpha, \beta, G)$ jointly with respect to (α, β) and G. Note that, for fixed (α, β), estimation of $G(t)$ is a special case of the biased sampling problem, discussed by Vardi (1985), Gill et al. (1988) and Qin (1993). Here, we adapt Anderson's (1972) approach by employing the Lagrange multiplier method to accomplish the maximisation. A different maximisation procedure was proposed by Prentice & Pyke (1979). From (1·2), we have

$$\mathscr{L}(\alpha, \beta, G) = \prod_{i=1}^{n_0} dG(x_i) \prod_{j=1}^{n_1} w(z_j)\, dG(z_j) = \left(\prod_{i=1}^{n} p_i \right) \left\{ \prod_{j=1}^{n_1} w(z_j) \right\}, \tag{2·1}$$

where $w(x) = \exp(\alpha + x\beta)$ and $p_i = dG(t_i)$ $(i = 1, 2, \ldots, n)$ are nonnegative jumps with total mass unity.

The first step is, for fixed (α, β), to maximise $\mathscr{L}$ with respect to p_i $(i = 1, \ldots, n)$ subject to contraints $\sum p_i = 1$, $p_i \geqslant 0$, $\sum p_i \{w(t_i) - 1\} = 0$, where the last constraint reflects the fact that $w(x)\, dG(x)$ is a distribution function. As in the approach of Qin & Lawless (1994), the maximum value of $\mathscr{L}$ is attained at $p_i = n_0^{-1}\{1 + \rho \exp(\alpha + t_i\beta)\}^{-1}$, where $\rho = n_1/n_0$. Therefore, ignoring a constant, the log-likelihood function is

$$l(\alpha, \beta) = \sum_{j=1}^{n_1} (\alpha + z_j\beta) - \sum_{i=1}^{n} \log\{1 + \rho \exp(\alpha + t_i\beta)\}. \tag{2·2}$$

Next we maximise l over (α, β). Let $(\tilde{\alpha}, \tilde{\beta})$ satisfy the following system of score equations:

$$\frac{\partial l(\alpha, \beta)}{\partial \alpha} = n_1 - \sum_{i=1}^{n} \frac{\rho \exp(\alpha + t_i\beta)}{1 + \rho \exp(\alpha + t_i\beta)} = 0, \quad \frac{\partial l(\alpha, \beta)}{\partial \beta} = \sum_{j=1}^{n_1} z_j - \sum_{i=1}^{n} \frac{t_i \rho \exp(\alpha + t_i\beta)}{1 + \rho \exp(\alpha + t_i\beta)} = 0. \tag{2·3}$$

Then we have

$$\tilde{p}_i = [n_0\{1 + \rho \exp(\tilde{\alpha} + t_i\tilde{\beta})\}]^{-1}. \tag{2·4}$$

Note that the score equations in (2·3) are identical to those of Prentice & Pyke (1979). The estimators $\tilde{p}_i$ can also be readily obtained from their paper.

On the basis of the $\tilde{p}_i$ in (2·4), we propose to estimate $G(t)$ by

$$\tilde{G}(t) = \sum_{i=1}^{n} \tilde{p}_i I(t_i \leqslant t) = \frac{1}{n_0} \sum_{i=1}^{n} \frac{I(t_i \leqslant t)}{1 + \rho \exp(\tilde{\alpha} + t_i\tilde{\beta})}. \tag{2·5}$$

If $\hat{G}(t)$ denotes the empirical distribution function based on control data $x_1, \ldots, x_{n_0}$, the

difference

$$\Delta(t) = n^{\frac{1}{2}} |\hat{G}(t) - \tilde{G}(t)|, \quad \Delta = \sup_{-\infty \leqslant t \leqslant \infty} \Delta(t) \tag{2·6}$$

may be used to measure the departure from the assumption of the logistic regression model (1·1). Note that, for two vectors $a = (a_1, \ldots, a_p)$ and $b = (b_1, \ldots, b_p)$, $a \leqslant b$ and $-\infty \leqslant a \leqslant \infty$ stand for, respectively, $a_i \leqslant b_i$ and $-\infty \leqslant a_i \leqslant \infty$ for $i = 1, \ldots, p$.

Remark 1. Similarly, $H(t) = F(t \mid y = 1)$ can be estimated by

$$\hat{H}(t) = \frac{1}{n_1} \sum_{j=1}^{n_1} I(z_j \leqslant t)$$

based on case data $z_1, \ldots, z_{n_1}$ and by

$$\tilde{H}(t) = \sum_{i=1}^{n} \tilde{p}_i \exp(\tilde{\alpha} + t_i \tilde{\beta}) I(t_i \leqslant t)$$

based on case-control data $t_1, \ldots, t_n$ under model (1·2). If

$$\Delta_1(t) = n^{\frac{1}{2}} |\tilde{H}(t) - \hat{H}(t)|, \quad \Delta_1 = \sup_{-\infty \leqslant t \leqslant \infty} \Delta_1(t),$$

then Δ_1 is an alternative test statistic. However, since $\Delta_1(t) = n_0 n_1^{-1} \Delta(t)$ and $\Delta_1 = n_0 n_1^{-1} \Delta$, there is a symmetry between the case and control designations for such a global test. Note that both Δ and Δ_1 reduce to the Kolmogorov–Smirnov two-sample statistic when $\alpha = \beta = 0$. Note also that some different distance measures for distribution functions other than Δ or Δ_1 can also be employed to test the validity of model (1·2). We will discuss them elsewhere.

Remark 2. In view of (2·2), we have

$$l(\alpha, \beta) = -\sum_{i=1}^{n} \log\{1 + \rho^{-1} \exp(-\alpha - t_i \beta)\} - \sum_{j=1}^{n_0} (\alpha + x_j \beta) - n \log \rho.$$

This can also be derived from the following model in which $(x_1, \ldots, x_{n_0})$ and $(z_1, \ldots, z_{n_1})$ are independent and

$x_1, \ldots, x_{n_0}$ are independent with density $g(x) = \exp(-\alpha - x\beta) h(x)$,

$z_1, \ldots, z_{n_1}$ are independent with density $h(x)$.

Remark 3. The test statistic Δ can also be applied to mixture sampling data in which a sample of $n = n_0 + n_1$ members is randomly selected from the whole population with both n_0 and n_1 being random (Day & Kerridge, 1967). Let (y_i, x_i) $(i = 1, 2, \ldots, n)$ be a random sample from the joint distribution of (y, x); then the likelihood has the form of

$$\mathscr{L} = \prod_{i=1}^{n} P(y_i \mid x_i) f(x_i) = \prod_{y_j = 1} \{\pi f(x_j \mid y = 1)\} \prod_{y_j = 0} \{(1 - \pi) f(x_j \mid y = 0)\},$$

where $\pi = P(y = 1)$.

Next we derive some asymptotic results. Let (α_0, β_0) be the true value of (α, β) under

model (1·2) and assume that $\rho = n_1/n_0$ remains fixed as $n \to \infty$. Then we have

$$-\frac{1}{n}\begin{pmatrix} \dfrac{\partial^2 l(\alpha_0, \beta_0)}{\partial \alpha^2} & \dfrac{\partial^2 l(\alpha_0, \beta_0)}{\partial \alpha\, \partial \beta^{\mathrm{T}}} \\ \dfrac{\partial^2 l(\alpha_0, \beta_0)}{\partial \beta\, \partial \alpha} & \dfrac{\partial^2 l(\alpha_0, \beta_0)}{\partial \beta\, \partial \beta^{\mathrm{T}}} \end{pmatrix}$$

$$\to \frac{\rho}{1+\rho}\begin{pmatrix} \displaystyle\int \frac{\exp(\alpha_0 + t\beta_0)}{1+\rho\exp(\alpha_0+t\beta_0)}\,dG(t) & \displaystyle\int \frac{t\exp(\alpha_0+t\beta_0)}{1+\rho\exp(\alpha_0+t\beta_0)}\,dG(t) \\ \displaystyle\int \frac{t^{\mathrm{T}}\exp(\alpha_0+t\beta_0)}{1+\rho\exp(\alpha_0+t\beta_0)}\,dG(t) & \displaystyle\int \frac{t^{\mathrm{T}}t\exp(\alpha_0+t\beta_0)}{1+\rho\exp(\alpha_0+t\beta_0)}\,dG(t) \end{pmatrix} = S$$

in probability, where $n_1/n = \rho/(1+\rho)$. Write

$$A_0(t) = \int_{-\infty}^{t} \frac{\exp(\alpha_0+\beta_0 y)}{1+\rho\exp(\alpha_0+\beta_0 y)}\,dG(y), \quad A_1(t) = \int_{-\infty}^{t} \frac{\exp(\alpha_0+\beta_0 y)}{1+\rho\exp(\alpha_0+\beta_0 y)}\,y\,dG(y),$$

$$A_2(t) = \int_{-\infty}^{t} \frac{\exp(\alpha_0+\beta_0 y)}{1+\rho\exp(\alpha_0+\beta_0 y)}\,y^{\mathrm{T}}y\,dG(y), \quad A_0 = A_0(\infty), \quad A_1 = A_1(\infty), \quad A_2 = A_2(\infty),$$

$$A = \begin{pmatrix} A_0 & A_1 \\ A_1^{\mathrm{T}} & A_2 \end{pmatrix}, \quad S = \frac{\rho}{1+\rho}A, \quad \Sigma = \frac{1+\rho}{\rho}\left[A^{-1} - \begin{pmatrix} 1+\rho & 0 \\ 0 & 0 \end{pmatrix}\right].$$

Then

$$n^{-\frac{1}{2}}\begin{pmatrix} \dfrac{\partial l(\alpha_0, \beta_0)}{\partial \alpha} \\ \dfrac{\partial l(\alpha_0, \beta_0)}{\partial \beta} \end{pmatrix} \to N(0, V)$$

in distribution, and

$$V = \frac{\rho}{1+\rho}A - \rho\begin{pmatrix} A_0 \\ A_1^{\mathrm{T}} \end{pmatrix}(A_0, A_1).$$

Now we can easily show the following.

LEMMA 1. *Under some regularity conditions, if model* (1·2) *is true then*

$$n^{\frac{1}{2}}\begin{pmatrix} \tilde{\alpha} - \alpha_0 \\ \tilde{\beta} - \beta_0 \end{pmatrix} \to N(0, S^{-1}VS^{-1}) = N(0, \Sigma)$$

in distribution.

This lemma matches Prentice & Pyke's (1979) results. In the following we consider the case of $p = 1$, though all the results can be naturally generalised to the case of $p > 1$.

THEOREM 1. *Under model* (1·2) *and suitable regularity conditions,*

$$\tilde{G}(t) - \hat{G}(t) = H_1(t) - \hat{G}(t) - H_2(t) + o_p(n^{-\frac{1}{2}}), \tag{2·7}$$

where

$$H_1(t) = \frac{1}{n_0}\sum_{i=1}^{n} \frac{I(t_i \leqslant t)}{1+\rho\exp(\alpha_0+\beta_0 t_i)}, \quad H_2(t) = \frac{\rho}{n}(A_0(t), A_1(t))S^{-1}\begin{pmatrix} \partial l(\alpha_0, \beta_0)/\partial\alpha \\ \partial l(\alpha_0, \beta_0)/\partial\beta \end{pmatrix}.$$

As a result, as $n \to \infty$, $n^{\frac{1}{2}}\{\tilde{G}(t) - \hat{G}(t)\}$ *converges weakly in* $D[-\infty, \infty]$ *to* $W(t)$, *a Gaussian process with mean* 0 *and covariance function specified by*

$$E\{W(s)W(t)\} = \rho(1+\rho)(A_0(s), A_1(s))\left\{\begin{pmatrix}1\\0\end{pmatrix} - A^{-1}\begin{pmatrix}A_0(t)\\A_1(t)\end{pmatrix}\right\}.$$

The proof of Theorem 1 will be given in the Appendix.

3. A BOOTSTRAP METHOD AND SOME NUMERICAL RESULTS

In this section we present a bootstrap method for finding critical values of Δ. If model (1·1) is valid, then, since α^* is not estimable in general on the basis of the case-control data $t_1, \ldots, t_n$, we can only generate bootstrap data from $d\tilde{G}(x)$ and $\exp(\tilde{\alpha} + x\tilde{\beta})\, d\tilde{G}(x)$, respectively. Specifically, let $x_1^*, x_2^*, \ldots, x_{n_0}^*$ be drawn independently from $d\tilde{G}(x)$ and independent of the x_i^*, and let $z_1^*, z_2^*, \ldots, z_{n_1}^*$ be drawn independently from $\exp(\tilde{\alpha} + x\tilde{\beta})\, d\tilde{G}(x)$. Note that some of the x_i^* could come from $z_1, \ldots, z_{n_1}$ and some of the z_j^* could be from $x_1, \ldots, x_{n_0}$. Let $t_1^*, \ldots, t_n^*$ be the combined bootstrap sample and let $(\tilde{\alpha}^*, \tilde{\beta}^*)$ be the solution to the system of equations in (2·3) with the t_i replaced by the t_i^*. Moreover, let $\hat{G}^*(t) = n_0^{-1}\sum_{i=1}^{n_0} I(x_i^* \leqslant t)$ and

$$\tilde{p}_i^* = \frac{1}{n_0}\frac{1}{1+\rho\exp(\tilde{\alpha}^* + t_i^*\tilde{\beta}^*)}, \quad \tilde{G}^*(t) = \frac{1}{n_0}\sum_{i=1}^{n}\frac{I(t_i^* \leqslant t)}{1+\rho\exp(\tilde{\alpha}^* + t_i^*\tilde{\beta}^*)}. \tag{3·1}$$

Then the corresponding bootstrap version of the test statistic Δ is

$$\Delta^*(t) = n^{\frac{1}{2}}|\hat{G}^*(t) - \tilde{G}^*(t)|, \quad \Delta^* = \sup_{-\infty \leqslant t \leqslant \infty} \Delta^*(t). \tag{3·2}$$

We propose to approximate the critical values of Δ by those of Δ^*.

In our simulation study we assume that $g(x)$ is the standard normal density function and $h(x)$ is the density function of a $N(\mu, \sigma^2)$ distribution. Then $g(x)$ and $h(x)$ are related by

$$h(x) = \exp(\alpha + x\beta + x^2\gamma)g(x), \tag{3·3}$$

where

$$\alpha = -\frac{1}{2}\left(\log\sigma^2 + \frac{\mu^2}{\sigma^2}\right), \quad \beta = \frac{\mu}{\sigma^2}, \quad \gamma = \frac{1}{2}\left(1 - \frac{1}{\sigma^2}\right). \tag{3·4}$$

It is clear that, if $\sigma = 1$, then $\gamma = 0$ and thus model (1·2) holds.

If model (3·3) is valid, then testing the validity of model (1·2) is equivalent to testing the null hypothesis $H_0: \gamma = 0$ under model (3·3). By analogy with the approach in § 2, we can maximise the semiparametric likelihood $\mathscr{L}(\alpha, \beta, \gamma, G)$ jointly with respect to (α, β, γ) and G under model (3·3), giving $(\tilde{\alpha}, \tilde{\beta}, \tilde{\gamma})$ and $\tilde{G}$. When $H_0: \gamma = 0$ is true under model (3·3), $n^{\frac{1}{2}}\tilde{\gamma} \to N(0, \sigma_\gamma^2)$ in distribution, where σ_γ^2 is the asymptotic variance of $\tilde{\gamma}$. If $\tilde{\sigma}_\gamma^2$ is the empirical version of σ_γ^2 on the basis of $\tilde{G}$ obtained as in (2·5), we can use the statistic

$$T = n^{\frac{1}{2}}\frac{\tilde{\gamma}}{\tilde{\sigma}_\gamma} \tag{3·5}$$

to test $H_0: \gamma = 0$ under model (3·3). Thus, both Δ and T can be used to test the validity of model (1·2). We would anticipate that T is better than Δ if model (3·3) is valid.

In our simulations, we considered $\gamma = 0, -0{\cdot}5, -1{\cdot}0$ and sample sizes of $n_0 = n_1 = 30$ and $n_0 = n_1 = 60$. In addition, we let $\beta = 0{\cdot}5$ be fixed. Note that $\alpha = -0{\cdot}125$, 0·2841, 0·5076 for $\gamma = 0, -0{\cdot}5, -1{\cdot}0$, according to (3·4). For each pair (n_0, n_1) and each value of γ, we generated 1000 independent sets of combined random samples from the $N(0, 1)$ and $N(\mu, \sigma^2)$ distributions, where μ and σ^2 were chosen to satisfy (3·4). Moreover, we generated 1000 independent combined bootstrap samples for each simulation. For significance levels 0·01, 0·05 and 0·1, the corresponding critical values of Δ were approximated by, respectively, the appropriate values in the ordered list of the 1000 bootstrap replications of Δ^*. The achieved significance levels and powers of the test statistic Δ can then be obtained from its position in the ordered list of the original combined random sample. The results in Table 1 show that the achieved significance levels of both Δ and T are quite close to the corresponding nominal significance levels and the powers of Δ and T are getting larger when γ is away from 0. As anticipated, the powers of T are all greater than those of Δ except for the case with $\gamma = -0{\cdot}5$ and $n_0 = n_1 = 30$.

Table 1. *Achieved significance levels and powers when* $h(x) = \exp(\alpha + x\beta + x^2\gamma)g(x)$ *with* $\beta = 0{\cdot}5$ *and* $g(x)$ *being the* $N(0, 1)$ *density function*

γ	Sample size	Levels	Δ	T	γ	Sample size	Levels	Δ	T
0·0	$n_0 = n_1 = 30$	0·10	0·147	0·102	0·0	$n_0 = n_1 = 60$	0·10	0·138	0·099
0·0	$n_0 = n_1 = 30$	0·05	0·075	0·035	0·0	$n_0 = n_1 = 60$	0·05	0·078	0·044
0·0	$n_0 = n_1 = 30$	0·01	0·016	0·006	0·0	$n_0 = n_1 = 60$	0·01	0·021	0·007
−0·5	$n_0 = n_1 = 30$	0·10	0·331	0·525	−0·5	$n_0 = n_1 = 60$	0·10	0·488	0·804
−0·5	$n_0 = n_1 = 30$	0·05	0·232	0·337	−0·5	$n_0 = n_1 = 60$	0·05	0·353	0·666
−0·5	$n_0 = n_1 = 30$	0·01	0·081	0·065	−0·5	$n_0 = n_1 = 60$	0·01	0·136	0·326
−1·0	$n_0 = n_1 = 30$	0·10	0·585	0·823	−1·0	$n_0 = n_1 = 60$	0·10	0·836	0·981
−1·0	$n_0 = n_1 = 30$	0·05	0·433	0·657	−1·0	$n_0 = n_1 = 60$	0·05	0·714	0·962
−1·0	$n_0 = n_1 = 30$	0·01	0·195	0·214	−1·0	$n_0 = n_1 = 60$	0·01	0·443	0·800

Next we consider two real datasets.

Example 1. Glovsky & Rigrodsky (1964) analysed and compared the developmental histories of mentally deficient children, some of whom had been diagnosed as asphasics. The subjects were 41 children enrolled in a speech therapy program at the Training School at Vineland, New Jersey. Twenty of the children had been diagnosed as asphasic during their early developmental years. The remaining subjects were a random sample of 21 mentally retarded children who were also enrolled in the speech therapy program. Let X represent 'Asphasic' and Z stand for 'Mentally retarded'. Then the social quotient scores made by these children on the Vineland Social Maturity Scale are given, respectively, by 56, 43, 30, 97, 67, 24, 76, 49, 46, 29, 46, 83, 93, 38, 25, 44, 66, 71, 54, 20, 25 for X and 90, 53 32, 44, 47, 42, 58, 16, 49, 54, 81, 59, 35, 81, 41, 24, 41, 61, 31, 20 for Z.

Daniel (1990, p. 336) also analysed this dataset by testing the null hypothesis $H_0: F_X = F_Z$ against the alternative hypothesis $H_A: F_X \neq F_Z$. Based on the Kolmogorov–Smirnov two-sample test statistic, Daniel concluded that we cannot reject H_0 in favour of H_A, and thus the two population distributions may not be different. Here, we analyse the data from a different perspective on the basis of model (1·2). The system of score equations in (2·3) yields $\tilde{\alpha} = 0{\cdot}39609$ and $\tilde{\beta} = -0{\cdot}00797$. Moreover, the proposed test statistic Δ in (2·6) is found to be $\Delta = 0{\cdot}566$. The observed P-value is 0·379. As a result, we cannot reject model (1·2).

Example 2. Hosmer & Lemeshow (1989, Ch. 1) used the logistic regression model (1·1) to analyse the relationship between age and the status of coronary heart disease based on 100 subjects participating in a study. The complete dataset is listed on page 3 in their book. Let x denote age and $y = 1$ or 0 represent the presence or absence of coronary heart disease. Since the data (y_i, x_i) $(i = 1, \ldots, 100)$ can be thought as being drawn independently and identically from the joint distribution of (y, x), Remark 3 in § 2 implies that we can make use of Δ in (2·6) and T in (3·5) to test the validity of model (1·1).

Under model (3·3) we find $T = 0{\cdot}25572$ and $(\tilde{\alpha}, \tilde{\beta}, \tilde{\gamma}) = (-3{\cdot}95890, 0{\cdot}06132, 0{\cdot}00055)$ with the observed P-value equal to 0·798, whereas under model (1·2) we have $\Delta = 0{\cdot}21994$ and $(\tilde{\alpha}, \tilde{\beta}) = (-5{\cdot}02760, 0{\cdot}11092)$ with the observed P-value equal to 0·979. Note that, since $n_0 = 57$ and $n_1 = 43$, α^* in model (1·1) can be estimated by $\tilde{\alpha}^* = -5{\cdot}0276 - \log(\frac{57}{43}) = -5{\cdot}3094$, which coincides with Hosmer & Lemeshow's result.

Figure 1 shows the curves of $\tilde{G}$ and $\hat{G}$ along with the curves $\tilde{H}$ and $\hat{H}$ based on this dataset. The curve of $\tilde{G}$ $(\tilde{H})$ bears a striking resemblance to that of $\hat{G}$ $(\hat{H})$.

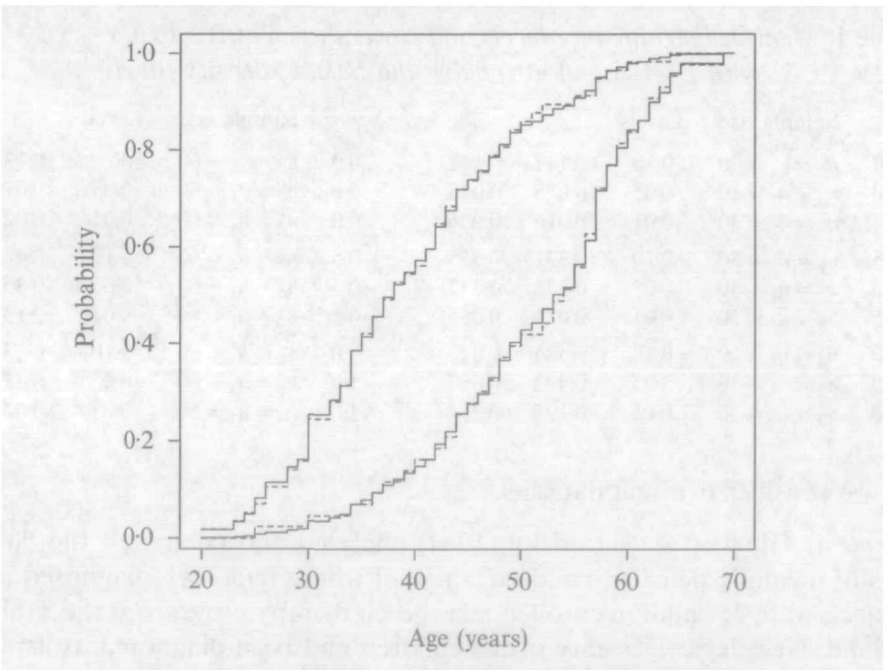

Fig. 1. Example 2: Estimated cumulative distribution functions $\tilde{G}$, $\hat{G}$, $\tilde{H}$ and $\hat{H}$. The solid curve and dashed curve on the upper left represent $\tilde{G}$ and $\hat{G}$, respectively, whereas the solid curve and dashed curve on the lower right stand for $\tilde{H}$ and $\hat{H}$, respectively.

4. DISCUSSION

The proposed procedure may be adapted to test the validity of other link functions for categorical data, such as model (3·3) or, based on case-control data, the probit model. Model (1·2) provides an alternative two-sample semiparametric model to the location-scale model and the Cox proportional hazards model.

If the logistic regression model (1·1) or its equivalent model (1·2) is valid, the next

problem is to test $\beta = 0$, that is, to test if x and y are independent or to test $f(x \mid y = 0) = f(x \mid y = 1)$. The score-based test statistic is

$$S = \frac{n_1}{n} \sum_{i=1}^{n} t_i - \sum_{j=1}^{n_1} z_j,$$

which is equivalent to $(\bar{x} - \bar{z})$. Furthermore, the bootstrap percentile method (Efron & Tibshirani, 1993, Ch. 13) can be used to construct confidence intervals for the odds ratio parameter β. Alternatively, as in Qin & Lawless (1994), we can construct confidence intervals for β based on empirical likelihood ratio statistics. Moreover, if we are interested in the underlying distribution function $G(x) = F(x \mid y = 0)$, it is possible to construct confidence bands for G based on a bootstrap approach.

Acknowledgement

We are grateful to the Editor, Professor Titterington, an associate editor and a referee for a number of helpful suggestions that have greatly improved our original submission, and to Professor Zhilian Ying for helpful comments.

Appendix

Proof of Theorem 1

Let

$$H_0(t) = \frac{1}{n_0} \sum_{i=1}^{n} \frac{\rho \exp(\alpha_0 + t_i\beta_0)}{\{1 + \rho \exp(\alpha_0 + t_i\beta_0)\}^2} I(t_i \leqslant t),$$
$$R_{1n}(t) = [H_0(t) - A_0(t), H_1(t) - A_1(t)] \begin{pmatrix} \tilde{\alpha} - \alpha_0 \\ \tilde{\beta} - \beta_0 \end{pmatrix}. \tag{A·1}$$

It can be shown by using a first-order Taylor expansion that

$$\tilde{G}(t) - \hat{G}(t) = H_1(t) - \hat{G}(t) - H_2(t) + R_{1n}(t) + R_{2n}(t),$$

where $\sup_{-\infty \leqslant t \leqslant \infty} |R_{2n}(t)| = o_p(n^{-\frac{1}{2}})$. In the proof of Lemma 1, we can obtain the asymptotic expression

$$\begin{pmatrix} \tilde{\alpha} - \alpha_0 \\ \tilde{\beta} - \beta_0 \end{pmatrix} = \frac{1}{n} S^{-1} \begin{pmatrix} \dfrac{\partial l(\alpha_0, \beta_0)}{\partial \alpha} \\ \dfrac{\partial l(\alpha_0, \beta_0)}{\partial \beta} \end{pmatrix} + o_p(n^{-\frac{1}{2}}),$$

which, along with (A·1), implies that

$$\sup_{-\infty \leqslant t \leqslant \infty} |R_{1n}(t)| = o_p(n^{-\frac{1}{2}}).$$

Let $R_n(t) = R_{1n}(t) + R_{2n}(t)$. Then (2·7) follows since $\sup_{-\infty \leqslant t \leqslant \infty} |R_n(t)| = o_p(n^{-\frac{1}{2}})$. For the proof of weak convergence of $n^{\frac{1}{2}}(\tilde{G} - \hat{G})$, according to (2·7), it suffices to show that

$$n^{\frac{1}{2}}\{H_1(t) - \hat{G}(t) - H_2(t)\} \to W(t)$$

weakly in $D[-\infty, \infty]$. It is easy to see that $E\{H_1(t) - \hat{G}(t) - H_2(t)\} = 0$. Moreover, very extensive algebra shows that

$$\operatorname{cov}(n^{\frac{1}{2}}\{H_1(s) - \hat{G}(s)\}, n^{\frac{1}{2}}H_2(t)) = \operatorname{cov}(n^{\frac{1}{2}}H_2(s), n^{\frac{1}{2}}H_2(t)),$$

and

$$\begin{aligned}
&\operatorname{cov}(n^{\frac{1}{2}}\{H_1(s)-\hat{G}(s)\}-n^{\frac{1}{2}}H_2(s), n^{\frac{1}{2}}\{H_1(t)-\hat{G}(t)\}-n^{\frac{1}{2}}H_2(t))\\
&\quad=\operatorname{cov}(n^{\frac{1}{2}}\{H_1(s)-\hat{G}(s)\}, n^{\frac{1}{2}}\{H_1(t)-\hat{G}(t)\})-\operatorname{cov}(n^{\frac{1}{2}}H_2(s), n^{\frac{1}{2}}H_2(t))\\
&\quad=\frac{n\rho}{n_0}A_0(s)-\frac{n^2\rho^2}{n_0n_1}A_0(s)A_0(t)-\rho(1+\rho)(A_0(s), A_1(s))A^{-1}\begin{pmatrix}A_0(t)\\A_1(t)\end{pmatrix}+\frac{n^2\rho^2}{n_0n_1}A_0(s)A_0(t)\\
&\quad=\rho(1+\rho)A_0(s)-\rho(1+\rho)(A_0(s), A_1(s))A^{-1}\begin{pmatrix}A_0(t)\\A_1(t)\end{pmatrix}\\
&\quad=\rho(1+\rho)(A_0(s), A_1(s))\left[\begin{pmatrix}1\\0\end{pmatrix}-A^{-1}\begin{pmatrix}A_0(t)\\A_1(t)\end{pmatrix}\right]=E\{W(s)W(t)\}.
\end{aligned}$$

This, together with the central limit theorem for sample means and the Cramer–Wold device, implies that the finite-dimensional distributions of $n^{\frac{1}{2}}\{H_1(t)-\hat{G}(t)-H_2(t)\}$ converge weakly to those of $W(t)$. Thus, in order to prove weak convergence of $n^{\frac{1}{2}}(\tilde{G}-\hat{G})$, it is enough to show that the process

$$\{n^{\frac{1}{2}}\{H_1(t)-\hat{G}(t)-H_2(t)\}, -\infty\leqslant t\leqslant\infty\}$$

is tight in $D[-\infty, \infty]$. However, this can be proved by employing the tightness criteria in Billingsley (1968, Ch. 3). The proof of Theorem 1 is completed.

REFERENCES

ANDERSON, J. A. (1972). Separate sample logistic discrimination. *Biometrika* **59**, 19–35.

AZZALINI, A., BOWMAN, A. & HARDLE, W. (1989). On the use of nonparametric regression for model checking. *Biometrika* **76**, 1–11.

BILLINGSLEY, P. (1968). *Convergence of Probability Measures*. New York: John Wiley.

BRESLOW, N. & DAY, N. E. (1980). *Statistical Methods in Cancer Research*, **1**, *The Analysis of Case-Control Studies*. Lyon: IARC.

DANIEL, W. W. (1990). *Applied Nonparametric Statistics*, 2nd ed. Boston: PKS-KENT.

DAY, N. E. & KERRIDGE, D. F. (1967). A general maximum likelihood discriminant. *Biometrics* **23**, 313–23.

EFRON, B. & TIBSHIRANI, R. (1993). *An Introduction to the Bootstrap*. London: Chapman & Hall.

FAREWELL, V. (1979). Some results on the estimation of logistic models based on retrospective data. *Biometrika* **66**, 27–32.

GILL, R. D., VARDI, Y. & WELLNER, J. A. (1988). Large sample theory of empirical distributions in biased sampling models. *Ann. Statist.* **16**, 1069–112.

GLOVSKY, L. & RIGRODSKY, S. (1964). A developmental analysis of mentally deficient children with early histories of aphasis. *Training School Bull.* **61**, 76–96.

HOSMER, D. J. & LEMESHOW, S. (1989). *Applied Logistic Regression*. New York: John Wiley.

LANDWEHR, J. M., PREGIBON, D. & SHOEMAKER, A. C. (1984). Graphical methods for assessing logistic regression models (with Discussion). *J. Am. Statist. Assoc.* **79**, 61–83.

PREGIBON, D. (1980). Goodness of link tests for generalized linear models. *Appl. Statist.* **29**, 15–24.

PRENTICE, R. L. & PYKE, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika* **66**, 403–11.

QIN, J. (1993). Empirical likelihood in biased sample problems. *Ann. Statist.* **21**, 1182–96.

QIN, J. & LAWLESS, J. F. (1994). Empirical likelihood and general estimating equations. *Ann. Statist.* **22**, 300–25.

VARDI, Y. (1982). Nonparametric estimation in presence of length bias. *Ann. Statist.* **10**, 616–20.

VARDI, Y. (1985). Empirical distribution in selection bias models. *Ann. Statist.* **13**, 178–203.

[*Received February* 1996. *Revised September* 1996]