Biometrika (1982), **69**, 1, *pp.* 123–36
Printed in Great Britain

Penalized maximum likelihood estimation in logistic regression and discrimination

BY J. A. ANDERSON
Department of Statistics, University of Newcastle upon Tyne
AND V. BLAIR
Statistical Unit, Christie Hospital, Manchester

SUMMARY

Maximum likelihood estimation of the parameters of the binary logistic regression model for $\mathrm{pr}(H \mid x)$ is discussed with separate discussion of sampling from (i) the conditional distribution of H given x, (ii) the joint distribution of H and x, and (iii) the conditional distribution of x given H. Difficulties associated with continuous x in the latter sampling scheme are discussed. To avoid these, penalized maximum likelihood estimates are introduced, which give estimates of the logistic parameters and a nonparametric spline estimate of the marginal distribution of x. Extensions to multinomial logistic regression are outlined.

Some key words: Binary logistic regression; Discrimination; Maximum likelihood estimation; Multinomial logistic regression; Nonparametric density estimation; Penalized maximum likelihood estimation; Spline function.

1. INTRODUCTION

Suppose that we are concerned with the relationship between a discrete variable H and a set of variables $x^{\mathrm{T}} = (x_1, \ldots, x_p)$, which may be continuous or discrete. We may be interested in predicting H from x or in making inferences about the dependence structure. In either case, the x_j will be referred to as regressor or predictor variables, whether nonrandom in a designed experiment or random in an observational study.

In many cases it is easy to model the conditional distribution of H given x, while the distribution of x given H is not so tractable. This is common in regression problems and presents no inferential difficulties when sample values are available from the distribution of H given x. This will be called x-conditional sampling. However, estimation and inference are not so straightforward when sample information is available only about the 'wrong' conditional distribution, that of x given H.

Anderson (1972) considered the above problem for binary or discrete H in the context of the logistic regression or discrimination model. The term separate sampling was used to indicate that separate samples had been taken from each of the distinct distributions of x given H. He showed that under separate sampling the standard estimates of the regression coefficients, defined in § 3·2, were ordinary maximum likelihood estimates for discrete x_j. Extrapolation to continuous x_j was given using a discretization argument. Prentice & Pyke (1979) claim that for continuous x_j, the standard estimators are maximum likelihood estimators. In § 3, we consider maximum likelihood estimation for binary H and give a new simple solution for discrete x, sampled separately. Our opinion

is that this approach is inadvisable for continuous x_j and instead we suggest using penalized maximum likelihood estimation, §4. The logistic parameter estimates are almost the same as the standard ones, §3·2. Farewell (1979) gave a conditional likelihood argument for separate sampling which leads to different estimates of the logistic parameters. There are no difficulties with continuous x_j but the method is computationally feasible only in special cases, such as one-to-k case-control matching.

In §5, H is permitted to be a general multinomial variable, while in §6 we examine the possibility of using penalized maximum likelihood methods as an alternative to conditional or partial likelihood methods when a full specification of the likelihood involves unknown functions in addition to the usual unknown parameters.

2. The logistic model

We wish to investigate the relationship between the binary variable H and $x^{\mathrm{T}} = (x_1, \ldots, x_p)$, which may be discrete or continuous. A convenient approach (Cox, 1970) is to take the logistic model for the conditional distribution of H given x:

$$\mathrm{pr}(H_1 \mid x) = e^z/(1+e^z), \quad \mathrm{pr}(H_2 \mid x) = 1/(1+e^z), \tag{1}$$

where $z = \alpha_0^* + \alpha^{\mathrm{T}} x$ and α_0^*, $\alpha^{\mathrm{T}} = (\alpha_1, \ldots, \alpha_p)$ are the unknown parameters, called logistic regression coefficients. This is a very natural approach when x-conditional sampling has been used; the corresponding maximum likelihood procedure is straightforward and given in §3·2.

'Mixture' sampling, where sample values are taken from the joint distribution of H and x, is often used. For example in a diagnostic study (Clayton, Anderson & McNicol, 1976) observations are taken on the predictor variables x and when available, the response variable H, disease category 1 or 2, is noted. Information about the conditional distribution of H given x is available by conditioning on x and hence inference about the model (1) is again straightforward. Section 3·3 deals with this.

Inferential difficulties do arise if mixture sampling is used, as is common in diagnostic studies (Anderson *et al.*, 1972), and in retrospective epidemiological investigations (Breslow & Powers, 1978; Prentice & Pyke, 1979). Here, x is sampled separately for $H = H_1$ and $H = H_2$, giving information about the likelihood of x given H_s, $L(x \mid H_s) = f_s(x)$, for $s = 1, 2$. We demonstrate that model (1) is still relevant. Under a wide variety of assumptions about the underlying distributions (Anderson, 1972, 1979, 1981a), the log likelihood ratio is linear in x

$$\log\{f_1(x)/f_2(x)\} = \alpha_0 + \alpha^{\mathrm{T}} x, \tag{2}$$

where α_0 and $\alpha^{\mathrm{T}} = (\alpha_1, \ldots, \alpha_p)$ are unknown parameters.

Our objective is then to make inferences about these parameters of the likelihood ratio. There is an obvious relationship between models (1) and (2). Under separate sampling, if the distributions $f_1(x)$ and $f_2(x)$ are mixed in the proportions π_1 and π_2, then the posterior probabilities $\mathrm{pr}(H_1 \mid x)$ and $\mathrm{pr}(H_2 \mid x)$ have the form specified in (1), provided that $\alpha_0^* = \alpha_0 + \log(\pi_1/\pi_2)$.

Thus the parameters α_0^* or α_0 and α arise quite naturally in investigating the relationship between H and x, no matter which sampling plan has been used. With separate sampling, a major objective is to make inferences about α_0 and α without making distributional assumptions other than those in (2). This is discussed in §3·4.

Note that the conditional distributions $f_1(x)$ and $f_2(x)$ can arise in mixture sampling by conditioning on $H = H_1$ or $H = H_2$. It will be useful in some contexts to estimate $f_1(x)$ and $f_2(x)$ for mixture or separate sampling, perhaps when considering the properties of a discriminant rule. This will be discussed in § 4.

3. Maximum likelihood estimation

3·1. *General*

The estimation problems arising in the three sampling plans are now considered in detail. It will be shown that the same function can be maximized to yield estimates of α_0^* or α_0 and α for all three plans with minor differences in interpretation. Recall that in separate sampling no assumptions have been made about the functional forms $f_1(.)$ and $f_2(.)$ so that their estimates are distribution-free, subject to the constraint (2).

3·2. *x-conditional sampling*

A total of $n(x)$ observations are made at x with $n_s(x)$ responses of H_s $(s = 1, 2)$. Note that $n(x)$ may be zero, indicating that no observations are taken at the x value. It follows from (1) that the likelihood is

$$L_c = \prod_{s=1}^{2} \prod_x \{\mathrm{pr}\,(H_s \mid x)\}^{n_s(x)} = \prod_x \left\{\frac{\exp(\alpha_0^* + \alpha^{\mathrm{T}} x)}{1+\exp(\alpha_0^* + \alpha^{\mathrm{T}} x)}\right\}^{n_1(x)} \left\{\frac{1}{1+\exp(\alpha_0^* + \alpha^{\mathrm{T}} x)}\right\}^{n_2(x)}. \quad (3)$$

Hence L_c is a function of the parameters α_0^* and α only and may be maximized iteratively, using any reliable optimization algorithm, to yield their maximum likelihood estimates. Note that α_0^* is estimable but α_0 is not. The functions $f_s(.)$ $(s = 1, 2)$ are not estimable. The above estimates of α_0^* and α are valid for any variables, continuous or discrete, satisfying (1) and since they are maximum likelihood estimates, their asymptotic properties are known. These are the 'standard' estimates of the coefficients referred to in § 1. It will be seen that they are appropriate in other sampling situations. The maximum of L_c is guaranteed to exist since L_c is bounded above by unity and below by zero. However, the maximum may be attained at nonunique points of the parameter space at infinity. This occurs when there is a hyperplane in the parameter space which completely separates the sample points from H_1 and H_2. This property of the sample space is easily detectable and is described in detail by Anderson (1972).

3·3. *Mixture sampling*

Sample values are available from the joint distribution of x and H. Suppose that $n_s(x)$ observations are noted at (x, H_s) for $s = 1, 2$ and, as before, let $n(x) = n_1(x) + n_2(x)$. Write the joint density $L(x, H) = \mathrm{pr}\,(H \mid x)\, L(x)$, where $L(x) = f(x)$, the marginal distribution of x. It follows that the likelihood of the observations is

$$L_m = L_c \prod_x \{f(x)\}^{n(x)}. \quad (4)$$

This likelihood contains as unknowns the parameters α_0^*, α and the function $f(.)$. The only constraint to be satisfied is that the function $f(.)$ is a density. This constraint depends in no way upon the values of the parameters α_0^* and α. Hence the maximum likelihood estimates of these parameters are obtained by maximizing L_c given in (3).

These estimates can also be justified as conditional maximum likelihood estimates since L_c is the likelihood obtained by conditioning on the $\{n(x)\}$.

Hence the procedure for obtaining the maximum likelihood estimates of α_0^* and α is the same for both x-conditional and mixture sampling; in both cases L_c is maximized. Additionally, for mixture sampling, the proportion of the population from H_s, π_s is estimable as $\hat{\pi}_s = n_s/n$ where $n_s = \Sigma n_s(x)$ $(s = 1, 2)$. This gives an estimate of α_0, $\hat{\alpha}_0 = \hat{\alpha}_0^* - \log(\hat{\pi}_1/\hat{\pi}_2)$. These estimates are equally valid for all variables, continuous or discrete, satisfying (1).

If all the variables are discrete, the pointwise maximum likelihood estimate of $f(x)$ exists and is $\hat{f}(x) = n(x)/n$. If any of the variables are continuous this is not the case, because the likelihood (4) is unbounded above.

3·4. *Separate sampling*

Here samples of fixed size n_s are taken from the conditional distributions $L(x \mid H_s)$ for $s = 1, 2$. As before, $n_s(x)$ observations from H_s are noted at x $(s = 1, 2)$, giving the likelihood as,

$$L_s = \prod_{s=1}^{2} \prod_x \{L(x \mid H_s)\}^{n_s(x)} = \prod_{s=1}^{2} \prod_x \{f_s(x)\}^{n_s(x)}. \tag{5}$$

In general, maximum likelihood estimation is much less straightforward here. When all the x variables are discrete, Anderson (1972) showed that the maximum likelihood estimates of α_0 and α could be obtained from the algorithm for maximizing L_c. One minor change is required; the parameter α_0^* in (3) is interpreted as being replaced by $\alpha_0 + \log(n_1/n_2)$. The algorithm is unchanged but note that α_0 is immediately estimable.

The reasoning of Anderson (1972) has been criticized by Farewell (1979) and Prentice & Pyke (1979) for lacking simplicity. The following shorter treatment may be preferred.

Using (1), the likelihood (5) can be written

$$L_s = \prod_x \{f_2(x)\}^{n(x)} \{\exp(\alpha_0 + \alpha^{\mathrm{T}} x)\}^{n_1(x)},$$

where $n(x) = n_1(x) + n_2(x)$. Now x is discrete and we suppose that it can take N distinct values. The parameters in L_s to be estimated are thus α_0, α and the N values of $f_2(x)$. The latter are regarded as multinomial probabilities. Thus $\log L_s$ is to be maximized with respect to these parameters and subject to the probability distribution constraints:

$$\sum_x f_1(x) = \sum_x \exp(\alpha_0 + \alpha^{\mathrm{T}} x) f_2(x) = 1, \quad \sum_x f_2(x) = 1.$$

When Lagrange multipliers are used to maximize $\log L_s$, the equation derived for $f_2(x)$ is

$$n(x)/f_2(x) + \lambda \exp(\alpha_0 + \alpha^{\mathrm{T}} x) + \mu = 0.$$

Multiplication by $f_2(x)$ and addition over x gives $n + \lambda + \mu = 0$, where $n = n_1 + n_2$. The equation for α_0 is

$$n_1 + \lambda \sum_x \exp(\alpha_0 + \alpha^{\mathrm{T}} x) f_2(x) = 0.$$

Hence $\lambda = -n_1$ and $\mu = -n_2$. This gives as maximum likelihood estimator of $f_2(x)$,

$$\hat{f}_2(x) = n(x)/\{n_2 + n_1 \exp(\alpha_0 + \alpha^{\mathrm{T}} x)\}.$$

Substitution of $\hat{f}_2(x)$ for $f_2(x)$ in L_s gives up to a constant multiplier

$$L_s(\hat{f}) = \prod_x \left\{ \frac{\exp(\alpha_0 + \alpha^{\mathrm{T}} x)}{n_2 + n_1 \exp(\alpha_0 + \alpha^{\mathrm{T}} x)} \right\}^{n_1(x)} \left\{ \frac{1}{n_2 + n_1 \exp(\alpha_0 + \alpha^{\mathrm{T}} x)} \right\}^{n_2(x)}.$$

It follows from Richards (1961) that the maximum likelihood estimates of α_0 and α are obtained by maximizing $L_s(\hat{f})$ and that their asymptotic covariance matrix is estimated from the Hessian of $\log\{L_s(\hat{f})\}$ at the maximum point. Further $L_s(\hat{f})$ is equal to L_c multiplied by a constant if α_0^* in L_c is replaced by $\alpha_0 + \log(n_1/n_2)$. For discrete x-variables, this completes the proof that maximum likelihood estimates of α_0 and α can be obtained by maximizing a modified L_c.

Anderson (1972) showed that the above procedure would also give approximate maximum likelihood estimates for continuous and/or discrete x variables. Prentice & Pyke (1979) showed from first principles that this estimation procedure is asymptotically unbiased with an asymptotic covariance as above. They also claim that this procedure gives maximum likelihood estimates. For continuous x, we do not believe that this is maximum likelihood estimation in the ordinary sense.

To see this, suppose that $\pi_1^\dagger$ and $\pi_2^\dagger$ are proportions $(\pi_1^\dagger + \pi_2^\dagger = 1)$ and let $f(x) = \pi_1^\dagger f_1(x) + \pi_2^\dagger f_2(x)$ be the mixture of the distributions from H_1 and H_2 in these proportions. Prentice & Pyke (1979) take $\pi_1^\dagger = n_1/n$ but we shall consider other values. It follows from (5) that

$$\begin{aligned} L_s &= \pi_1^{\dagger -n_1} \pi_2^{\dagger -n_2} \left[\prod_{s=1}^{2} \prod_x \left\{ \frac{\pi_s^\dagger f_s(x)}{\pi_1^\dagger f_1(x) + \pi_2^\dagger f_2(x)} \right\}^{n_s(x)} \right] \prod_x \{\pi_1^\dagger f_1(x) + \pi_2^\dagger f_2(x)\}^{n(x)} \\ &= \pi_1^{\dagger -n_1} \pi_2^{\dagger -n_2} \left[\prod_{s=1}^{2} \prod_x \{P_s^\dagger(x)\}^{n_s(x)} \right] \prod_x \{f(x)\}^{n(x)}, \end{aligned} \tag{6}$$

where $P_1^\dagger(x) = e^{z^\dagger}/(1+e^{z^\dagger})$, $P_2^\dagger(x) = 1/(1+e^{z^\dagger})$, $z^\dagger = \alpha_0 + \log(\pi_1^\dagger/\pi_2^\dagger) + \alpha^{\mathrm{T}} x$. Hence (6) can be written as

$$L_s = \pi_1^{\dagger -n_1} \pi_2^{\dagger -n_2} L_c^\dagger \prod_x \{f(x)\}^{n(x)}, \tag{7}$$

where $L_c^\dagger = L_c$ as given in (3) with α_0^* replaced by $\alpha_0^\dagger$, where

$$\alpha_0^\dagger = \alpha_0 + \log(\pi_1^\dagger/\pi_2^\dagger). \tag{8}$$

The likelihood L_s in (10) thus contains as unknowns $\alpha_0^\dagger$, α and the function $f(\,.\,)$. These must satisfy the conditions that $f_s(\,.\,)$ is a density $(s = 1, 2)$, that is

$$\int f_s(x)\,dx = \frac{1}{\pi_s^\dagger} \int P_s^\dagger(x)\, f(x)\,dx = 1 \quad (s = 1, 2), \quad f(x) \geqslant 0 \tag{9}$$

for all x.

First we consider the case where all the x variables are continuous so that the density $f(x)$ is continuous with respect to Lebesgue measure in the usual way. Consideration of the likelihood L_s in (7) indicates that, regarded as a functional of $f(\,.\,)$, it is not bounded above. This follows from the type of arguments employed by Good & Gaskins (1971). Thus in one dimension, take $f(x)$ as the arithmetic mean of the n delta functions located at the n sample points and take $P_s^\dagger(x) = \pi_s^\dagger$ $(s = 1, 2)$. This gives an infinite value for L_s while satisfying the constraints (9). A sequence of normal densities with variances

tending to zero could also illustrate the point. The same considerations apply in higher dimensions. Thus, unlike Prentice & Pyke (1979), we conclude that maximum likelihood estimation is not appropriate for maximizing L_s in (7) if any of the $\{x_j\}$ are continuous. Instead, we suggest using penalized maximum likelihood estimation for continuous data in separate sampling and for mixture sampling when an estimate of the mixture density is also required. We shall give a full solution for the case $p = 1$ and show that the estimates of α_0 and α can still be found using the algorithm for maximizing L_c. In higher dimensions the problem is more difficult but we show that the same sort of solution is obtained although we are unable to produce explicit solutions for $f_s(x)$.

4. Maximum penalized likelihood estimation for continuous densities

4·1. *General considerations*

It has been shown that maximum likelihood estimation leads to rough estimates of density functions for continuous variables under the minimum restrictions imposed by (2). However, the density functions of most naturally occurring random variables are quite smooth and hence it seems reasonable to demand that our estimates should also be smooth in some sense. To achieve this, we suggest using Good & Gaskins's (1971) penalized maximum likelihood procedure. This was developed to estimate a single continuous density function, $g(.)$ say. The method involves subtracting from the log likelihood a strictly positive term that provides a measure of the roughness of the density function. The resulting penalized log likelihood is maximized to obtain a smooth estimate of the continuous density function $g(.)$. Good & Gaskins (1971) show that under fairly weak conditions on the density $g(.)$ and the penalty function that the penalized maximum likelihood estimate $\tilde{g}(.)$ of $g(.)$ is consistent in the sense that

$$\mathrm{pr}\left[\left|\int_a^b \{\tilde{g}(x)-g(x)\}\right| dx < \varepsilon\right] \to 1$$

as the number of observations tends to infinity for all $\varepsilon > 0$ and for all a, b such that $a < b$. To our knowledge, no other consistency or asymptotic properties have been derived for maximum penalized likelihood estimation but we believe that it will prove to be asymptotically efficient. One penalty function proposed by Good & Gaskins (1971) for estimating a continuous univariate density function $g(.)$ is

$$K \int \{g'(x)\}^2/g(x)\, dx, \tag{10}$$

where $K > 0$. In principle, the methods of this section are applicable to any choice of the penalty function. The function (10) prevents the derivative of the density estimate from becoming too large, so that the estimate will be fairly smooth and bounded. If the penalty function is to be finite for the true density the first derivative of the square root of this density must be of integrable square. Finiteness of the penalty function for the true density is required by Good & Gaskins's (1971) consistency proof. Henceforth it will be assumed that all density functions of univariate continuous variables satisfy this fairly weak condition. The constant K determines how smooth the estimate will be. Some method for selecting a value of K has to be used. Wahba & Wold (1975) suggest the use of cross-validation techniques in a similar context but this is costly in computer time. To

date, probably the most frequently used method is to try several values of K and select one for which the small scale variation of the estimate disappears but for which the large scale variation is still present.

Sections 4·2 and 4·3 deal with the application of this technique to the one-dimensional case for mixture and separate sampling respectively. Suitable penalty functions for two or more dimensions can be constructed but so far they have not proved to be mathematically tractable. An outline of the technique for $p \geqslant 2$ will be given in §§ 4·4 and 4·5.

4·2. *Mixture sampling of one continuous variable*

In the notation of § 3·3 and with the penalty function defined in (10), the penalized log likelihood for mixture sampling of one continuous variable is, by (4),

$$\log L_m - K\int_I [\{f'(x)\}^2/f(x)]\,dx = \log L_c + \textstyle\sum_x n(x)\log f(x) - K\int_I [\{f'(x)\}^2/f(x)]\,dx, \quad (11)$$

where I is an interval on the real line, finite or infinite, such that $f(x) \equiv 0$ for all $x \notin I$. The problem is to maximize (11) with respect to α_0^*, α and $f(\,.\,)$ subject to the constraint that $f(\,.\,)$ is a density function, that is

$$\int_I f(x)\,dx = 1, \quad f(x) \geqslant 0 \quad (x \in I). \quad (12)$$

As in § 3·3, it can be seen that the estimates of α_0^* and α are obtained by maximizing L_c. Hence this method produces the same estimates for α_0^* and α as did maximum likelihood estimation. The estimate of $f(\,.\,)$ is obtained by maximizing

$$\textstyle\sum_x n(x)\log f(x) - K\int_I [\{f'(x)\}^2/f(x)]\,dx \quad (13)$$

subject to the constraints (12). This gives a spline function as detailed in the Appendix.

Having found the penalized maximum likelihood estimates $\tilde{\alpha}_0^*$, $\tilde{\alpha}$ and $\tilde{f}(\,.\,)$, we obtain the corresponding estimates of π_1, π_2, α_0, $f_1(\,.\,)$ and $f_2(\,.\,)$ as

$$\tilde{\pi}_1 = \int_I \tilde{P}(H_1\,|\,x)\tilde{f}(x)\,dx, \quad \tilde{\pi}_2 = 1-\tilde{\pi}_1, \quad \tilde{\alpha}_0 = \tilde{\alpha}_0^* - \log(\tilde{\pi}_1/\tilde{\pi}_2), \quad (14)$$

$$\tilde{f}_1(x) = \tilde{\pi}_1^{-1}\tilde{P}(H_1\,|\,x)\tilde{f}(x), \quad \tilde{f}_2(x) = \tilde{\pi}_2^{-1}\tilde{P}(H_2\,|\,x)\tilde{f}(x), \quad (15)$$

where

$$\tilde{P}(H_1\,|\,x) = e^{\tilde{z}}/(1+e^{\tilde{z}}), \quad \tilde{P}(H_2\,|\,x) = 1-\tilde{P}(H_1\,|\,x), \quad \tilde{z} = \tilde{\alpha}_0^* + \tilde{\alpha}^{\mathrm{T}}x. \quad (16)$$

The use of the penalty function has destroyed some of the symmetry of the problem in the sense that the estimator $\tilde{\pi}_1$ of π_1 is not generally the proportion of observations from H_1 and hence the estimate of α_0 is not the same as that given in § 3·3.

Note that the use of any other penalty function in place of (10) will give the same form for the above results, but the solution for $\tilde{f}(\,.\,)$ will not be that of the Appendix. If an exact solution for $\tilde{f}(\,.\,)$ is not available, (14) and (15) will not be exactly calculable. See § 4·4 for more discussion of this point.

4·3. *Separate sampling for one continuous variable*

Here there is no obvious choice for the density function to be penalized in (10). It seems reasonable to demand that the estimation system should be invariant under the labelling of the two populations as H_1 and H_2. This immediately rules out the choice of either $f_1(.)$ or $f_2(.)$ as the function to be penalized, if we use the notation of § 3·4. Instead, we suggest penalizing the mixture density $f(x) = \pi_1^\dagger f_1(x) + \pi_2^\dagger f_2(x)$, where the choice of $\pi_1^\dagger$ to give labelling invariance, will be dealt with later. The resulting penalized log likelihood is, from (7) and (10),

$$\log L_s - K\int_I [\{f'(x)\}^2/f(x)]\,dx = -n_1 \log \pi_1^\dagger - n_2 \log \pi_2^\dagger + \log L_c^\dagger + \sum_x n(x) \log f(x) - K\int_I [\{f'(x)\}^2/f(x)]\,dx, \quad (17)$$

where I is as defined in § 4·2. This has to be maximized, subject to the constraints (9), to obtain point estimates of $\alpha_0^\dagger$, defined in (8), and α and a smooth estimate of $f(.)$. From these, estimates of α_0, $f_1(.)$ and $f_2(.)$ can be calculated. The problem is now almost identical to that under the mixture sampling plan. Apart from the addition of a known constant, the penalized log likelihoods (11) and (17) of the two problems are identical if α_0^* is identified with $\alpha_0^\dagger$. The constraints (9) that have to be satisfied are equivalent to the constraints (12) of the mixture sampling problem plus the additional constraint

$$\int P_1^\dagger(x)\, f(x)\, dx = \pi_1^\dagger. \quad (18)$$

Because this extra constraint has to be satisfied the estimation of $\alpha_0^\dagger$ and α no longer separates from that of $f(.)$ for arbitrary $\pi_1^\dagger$ and $\pi_2^\dagger$.

The proportions $\pi_1^\dagger$ and $\pi_2^\dagger$ have to be chosen to give invariance under the labelling of H_1 and H_2. An obvious choice for $\pi_1^\dagger$ would seem to be n_1/n but the only solution found has $K = 0$ which gives the very rough density estimates that have already been rejected. Alternatively, if the same data had been obtained by mixture sampling it would have contained exactly the same amount of information about the two densities $f_1(.)$ and $f_2(.)$; the extra information in the mixture sample would just relate to the unknown mixing proportions. Hence it is reasonable to use a value of $\pi_1^\dagger$ which ensures that the estimates of $f_1(.)$ and $f_2(.)$ are the same as those that would have been obtained had the observations been drawn from a mixture of the two populations. There is only one value for $\pi_1^\dagger$ for which this is the case; this is the estimate $\tilde{\pi}_1$ of the mixture proportion, defined in (14), that would have been obtained if a mixture sampling scheme had been used to collect the data. It is fairly straightforward to see that this choice gives labelling invariance.

The maximization of (17) with respect to α_0, α and $f(.)$, subject to the constraints (12) and (18) is greatly simplified by putting $\pi_1^\dagger = \tilde{\pi}_1$. The penalized maximum likelihood estimates are given by the following procedure: (i) maximize $L_c^\dagger$ with respect to $\alpha_0^\dagger$ and α, (ii) maximize (13) with respect to $f(.)$ subject to constraint (12). The sum of $L_c^\dagger$ and (13) in (17) is maximized as the two parts have been separately maximized. Condition (12) has been satisfied and condition (18) is satisfied because we have taken $\pi_1^\dagger = \tilde{\pi}_1$. Note that equation (18) gives the value of $\tilde{\pi}_1$, when we substitute the estimates of α_0, α and $f(.)$.

The corresponding estimates of α_0, $f_1(.)$ and $f_2(.)$ are

$$\tilde{\alpha}_0 = \tilde{\alpha}_0^\dagger - \log(\tilde{\pi}_1/\tilde{\pi}_2), \quad \tilde{f}_1(x) = \tilde{P}_1^\dagger(x)\tilde{f}(x)/\tilde{\pi}_1, \quad \tilde{f}_2(x) = \tilde{P}_2^\dagger(x)\tilde{f}(x)/\tilde{\pi}_2, \tag{19}$$

where

$$\tilde{P}_1^\dagger(x) = e^{\tilde{z}^\dagger}/(1+e^{\tilde{z}^\dagger}), \quad \tilde{P}_2^\dagger(x) = 1/(1+e^{\tilde{z}^\dagger}), \quad \tilde{z}^\dagger = \tilde{\alpha}_0^\dagger + \tilde{\alpha}^{\mathrm{T}} x. \tag{20}$$

The estimate of α is that of Anderson (1972) and Prentice & Pyke (1979) for which they have presented the asymptotic properties. Our estimates of α_0 differs from that of Anderson (1972) and Prentice & Pyke (1979) by $\log(n_1/n_2) - \log(\tilde{\pi}_1/\tilde{\pi}_2)$. If, under the mixture sampling plan, $\tilde{\pi}_1$, defined in (14), is a consistent estimator of the mixing proportion π_1, then in some sense $\log(n_1/n_2) - \log(\tilde{\pi}_1/\tilde{\pi}_2) \to 0$ as $n \to \infty$; in which case, for a large enough sample the probability that the two estimates differ greatly will be small.

In practice, we may want to estimate posterior probabilities for discrimination between the populations H_1 and H_2, where they occur in the proportions π_1 and π_2. This implies that an estimate of $\alpha_0 + \log(\pi_1/\pi_2)$ is required and an obvious choice is $\tilde{\alpha}_0 + \log(\pi_1/\pi_2)$. Provided that we keep the same function $f(x) = \pi_1^\dagger f_1(x) + \pi_2^\dagger f_2(x)$ in (21) with $\pi_1^\dagger = \tilde{\pi}_1$, $\tilde{\alpha}_0 + \log(\pi_1/\pi_2)$ is a penalized maximum likelihood estimate. However, if we take the more natural approach and change $f(x)$ by taking $\pi^\dagger = \pi_1$, to correspond to the mixture of discriminant interest, the optimization problem is difficult and an explicit solution is not available. As in §4·2, similar results are available for penalty functions other than (10).

4·4. *Multivariate continuous data*

Suppose that a suitable penalty function, $\Phi(f)$ say, for multivariate continuous data has been constructed so that the solution to the problem of maximizing

$$\Sigma\, n(x) \log f(x) - K\Phi(f), \tag{21}$$

where $K > 0$, subject to the constraints

$$\int_I f(x)\, dx = 1, \quad f(x) \geqslant 0 \quad (x \in I), \tag{22}$$

is smooth in some sense and bounded. Good & Gaskins (1971) discuss approximations to solutions of this problem in terms of Hermite functions of several variables and generalized Hermite functions. We have not yet been able to find a multivariate version of the penalty function that provides an explicit smooth density estimate. The problems associated with the multivariate case will be presented elsewhere. For the moment, suppose that $\tilde{f}(.)$ is a multivariate penalized density estimate, either exact or approximate, obtained by maximizing (21) subject to (22).

In the notation of §3·3, it follows from equation (4) for the mixture sampling scheme that the penalized log likelihood

$$\log L_m - K\Phi(f) = \log L_c + \textstyle\sum_x n(x) \log f(x) - K\Phi(f) \tag{23}$$

can be formed. This contains as unknowns α_0^*, α and $f(.)$ and can be maximized, subject to the constraint that $f(.)$ is a density function, to obtain estimates of these quantities. Thus, it can be seen that the estimates of α_0^* and α are obtained by maximizing L_c and that the estimate of $f(.)$ is $\tilde{f}(.)$, introduced above. Estimates of π_1, π_2, α_0, $f_1(.)$ and

$f_2(.)$ are given by the multivariate versions of (14) and (15). Note that the explicit form of $\tilde{f}(.)$ is required for the estimate, $\tilde{\pi}_1$, of π_1 and hence for the estimate, $\tilde{\alpha}_0$, of α_0. If $\tilde{f}(.)$ is not available, π_1 may be estimated by n_1/n following Anderson (1972) and Prentice & Pyke (1979). As noted at the end of §4·3, as the sample size increases, the difference between $\tilde{\pi}_1$ and n_1/n should tend to zero in some sense. This question does not arise in contexts like discrimination, where the prime interest is in estimating α_0^*. The usual asymptotic properties of the estimates $\tilde{\alpha}_0^*$ and $\tilde{\alpha}$ follow from the argument given by Cox (1975). The asymptotic properties of $\tilde{\alpha}_0$, $\tilde{\pi}_1$, $\tilde{\pi}_2$, $\tilde{f}_1(.)$ and $\tilde{f}_2(.)$ depend upon those of the estimate $\tilde{f}(.)$ of the mixture density $f(.)$.

The treatment for separate sampling of multivariate continuous data follows along similar lines. As in §4·2, a penalty function depending on a known mixture density $f(x) = \pi_1^\dagger f_1(x) + \pi_2^\dagger f_2(x)$ is subtracted from the log likelihood to give the penalized log likelihood, by (7),

$$-n_1 \log \pi_1^\dagger - n_2 \log \pi_2^\dagger + \log L_c^\dagger + \textstyle\sum_x n(x) \log f(x) - K\Phi(f), \tag{24}$$

which depends on the unknowns $\alpha_0^\dagger = \alpha_0 + \log(\pi_1^\dagger/\pi_2^\dagger)$, α and $f(.)$. If we follow the argument of §4·3, this has to be maximized subject to constraint (22) and the additional constraint

$$\int_I P_1^\dagger(x) f(x)\, dx = \pi_1^\dagger. \tag{25}$$

If $\pi_1^\dagger$ is chosen equal to $\tilde{\pi}_1$ as in §4·2, then it is obvious that the estimates of $\alpha_0^\dagger = \alpha_0 + \log(\tilde{\pi}_1/\tilde{\pi}_2)$ and α are obtained by maximizing $L_c^\dagger$ and the estimate of $f(.)$ by maximizing (21) subject to (22). The corresponding estimates of α_0, $f_1(.)$ and $f_2(.)$ are given by (20) with the multivariate version of $\tilde{f}_1(x)$ and $\tilde{f}_2(x)$. The asymptotic properties of $\tilde{\alpha}$ are given by Prentice & Pyke (1979), those of $\tilde{\alpha}_0$, $\tilde{f}_1(.)$ and $\tilde{f}_2(.)$ depend on the asymptotic properties of $\tilde{f}(.)$. As in §4·3, the value of $\tilde{\pi}_1$ is found by substituting estimates of α_0, α and $f(.)$ into (18). Further, $\tilde{\pi}_1$ must be approximated by n_1/n if an explicit form for $f(.)$ is not available. This is now relevant for discrimination as the optimum allocation rule depends on $\tilde{\alpha}_0^\dagger - \log(\tilde{\pi}_1/\tilde{\pi}_2) + \log(\pi_1/\pi_2)$, when discriminating between the two populations mixed in the ratio $\pi_1 : \pi_2$.

4·5. *Combined continuous/discrete data*

Suppose that $x^{\mathrm{T}} = (x_1^{\mathrm{T}}, x_2^{\mathrm{T}})$, where each component of x_1 is continuous and each component of x_2 is discrete.

First consider mixture sampling. Write the mixture density $\pi_1 f_1(x) + \pi_2 f_2(x)$ as $f(x_1) g(x_2 \mid x_1)$, where $f(x_1)$ is the mixture density of x_1 and $g(x_2 \mid x_1)$ is the mixture density of x_2 given x_1. The log likelihood is, by (4),

$$\log L_c + \textstyle\sum_x n(x) \log \{f(x_1)\, g(x_2 \mid x_1)\}, \tag{26}$$

where $n(x) = n(x_1, x_2)$ is the number of sample points from H_1 and H_2 at $x^{\mathrm{T}} = (x_1^{\mathrm{T}}, x_2^{\mathrm{T}})$. Expression (26) depends on α_0^*, α, $f(.)$ and $g(.)$. To estimate these quantities a suitable penalized log likelihood to maximize is

$$\log L_c + \textstyle\sum_x n(x) \log g(x_2 \mid x_1) + \sum_x n(x) \log f(x_1) - K\Phi(f), \tag{27}$$

subject to the constraints

$$\int_I f(x_1)\,dx_1 = 1, \quad \textstyle\sum_{x_2} g(x_2 \mid x_1) = 1 \quad (-\infty < x_1 < \infty),$$
$$f(x_1) \geqslant 0 \quad (x_1 \in I), \quad g(x_2 \mid x_1) \geqslant 0 \quad (-\infty < x_1, x_2 < \infty). \tag{28}$$

Hence the estimates of α_0^* and α are obtained by maximizing L_c and the estimate of $f(\,.\,)$ by maximizing (21) subject to (22). The estimate of $g(\,.\mid.\,)$ maximizes

$$\textstyle\sum_x n(x) \log g(x_2 \mid x_1), \tag{29}$$

subject to

$$\textstyle\sum_{x_2} g(x_2 \mid x_1) = 1, \quad g(x_2 \mid x_1) \geqslant 0. \tag{30}$$

Thus the estimate of $g(x_2 \mid x_1)$ is $\tilde{g}(x_2 \mid x_1) = n(x_1, x_2)/\sum_{x_2} n(x_1, x_2)$. Recall that $n(x) = n(x_1, x_2)$. Estimates of π_1, π_2 and the densities in each population at (x_1, x_2) can be obtained from analogues of equations (14) and (15). These involve the explicit form of the estimate of $f(\,.\,)$, which may be unknown for multivariate x_1, as in the previous section. At worst, π_1 may be estimated by n_1/n so that an estimate of α_0 based on $\tilde{\alpha}_0$ in (14) is always available.

The separate sampling problem for combined continuous/discrete data follows along similar lines. Once again, a known mixture density is formed $\pi_1^\dagger f_1(x_1) + \pi_2^\dagger f_2(x_2) = f(x_1)\, g(x_2 \mid x_1)$, where $f(\,.\,)$ and $g(\,.\mid.\,)$ have the same interpretation as above. By (7), this leads to the penalized log likelihood

$$-n_1 \log \pi_1^\dagger - n_2 \log \pi_2^\dagger + \log L_c^\dagger + \textstyle\sum_x n(x) \log g(x_2 \mid x_1) + \sum_x n(x) \log f(x_1) - K\Phi(f) \tag{31}$$

depending on the unknowns $\alpha_0^\dagger = \alpha_0 + \log(\pi_1^\dagger/\pi_2^\dagger)$, α, $f(\,.\,)$ and $g(\,.\mid.\,)$ which have to satisfy the constraints (29) plus the additional constraint

$$\pi_1^\dagger = \int_I f(x_1) \{\textstyle\sum_{x_2} P^\dagger(x)\, g(x_2 \mid x_1)\}\, dx_1. \tag{32}$$

If $\pi_1^\dagger$ is chosen equal to $\tilde{\pi}_1$, as in §4·3, the estimates of $\alpha_0^\dagger$ and α are found by maximizing $L_c^\dagger$, the estimate of $f(\,.\,)$ by maximizing (21) subject to (22) and the estimate of $g(\,.\mid.\,)$ by maximizing (29) subject to (30). As in §4·3, the value of $\tilde{\pi}_1$ is found by substituting estimated values into (18) and then the estimate of α_0 is given by (19). The comments in §§4·3 and 4·4 about approximating $\tilde{\pi}_1$ by n_1/n also apply here, but recall that if x_1 is one dimensional, $\tilde{\pi}_1$ can and should be found exactly. The asymptotic properties of the estimates are similar to those of previous sections.

5. Multinomial values for H

We have been concerned so far with models for the relationship between a binary variable H and $x^{\mathrm{T}} = (x_1, \ldots, x_p)$. These methods generalize immediately to the case where H is a multinomial variable taking values $H_1, \ldots, H_k$. The multinomial logistic model corresponding to (1) is

$$\mathrm{pr}(H_s \mid x) = e^{z_s} \Big/ \left(1 + \sum_{t=1}^{k-1} e^{z_t}\right) \quad (s = 1, \ldots, k-1), \quad \mathrm{pr}(H_k \mid x) = 1 \Big/ \left(1 + \sum_{t=1}^{k-1} e^{z_t}\right),$$

where $z_s = \alpha_{0s}^* + \alpha_s^{\mathrm{T}} x$ and α_s is a p-component vector of parameters ($s = 1, \ldots, k-1$). An expression for $L_c^{(k)}$, similar to L_c in (3), follows for x-conditional and mixture sampling. The usual maximum likelihood estimates for the parameters α_{0s} and α_s ($s = 1, \ldots, k-1$) are obtained by maximizing $L_c^{(k)}$. Again, the separate sampling case is more difficult, but Anderson (1972) gave an approach for discrete x, which can be simplified as in § 3·4 when using the likelihood ratio model equivalent to (2).

For continuous x, penalized maximum likelihood estimates of α_{0s}^*, α_s ($s = 1, \ldots, k-1$) and $f(.)$ for mixture sampling are obtained by substituting $L_c^{(k)}$ in (11) or (24) as appropriate and maximizing as in § 4. The estimates of α_{0s}^* and α_s are obtained by maximizing $L_c^{(k)}$. Continuous variables x in separate sampling are dealt with by introducing $L_c^{\dagger(k)}$, analogous to $L_c^{\dagger}$ in 3·4, which replaces $L_c^{\dagger}$ in (17) or (24).

6. DISCUSSION

For continuous regressor variables, we have shown that a penalized maximum likelihood approach to the estimation of the parameters α_0 and α in (2) yields the same estimators of α as the standard method given in § 3·2. The penalized estimator of α_0 is different from its standard estimator but they will be approximately equal in large samples. It may seem paradoxical that there should be two estimators, $\hat{\alpha}_0$ and $\tilde{\alpha}_0$, for α_0, particularly for mixture sampling. However, the two estimators are obtained under different levels of assumptions about the underlying distributions. To obtain $\tilde{\alpha}_0$, we have insisted on a smooth estimate of the underlying smooth density function. No such constraint was imposed to derive $\hat{\alpha}_0$ and hence the corresponding estimate of the density is very rough; it is zero everywhere except at those points which occurred in the sample. This is an odd estimate of a smooth density function.

In many statistical problems, the full likelihood involves unknown parameters and an unknown smooth, continuous function. To estimate the parameter without making assumptions about the unknown function requires ingenuity and may involve conditional, marginal or partial likelihood estimation. For example, see the papers by Cox (1972, 1975) and Kalbfleisch & Prentice (1973) on a regression model for life tables.

When an estimate of the unknown function is also required, there is even less agreement about a general approach. Usually the continuous function is approximated by a step-function which can be estimated simultaneously with the parameters or conditionally given the parameter estimates. However, penalized maximum likelihood gives a method of simultaneously estimating the function and the parameters without making arbitrary approximations to the function. This paper gives examples of this in § 4, but the approach may not always be computationally feasible. A simpler method is to take a penalized estimate of the function, conditional on the parameter estimates. Anderson & Senthilselvan (1980) used this method to derive a smooth estimate of the hazard function in Cox's (1972) model. This conditioning may involve some information loss. If so, an iterative scheme could be set up alternatively conditioning on the parameters and estimating the functions and *vice versa*. These methods are also applicable to generalizations of the problems of this paper where the likelihood ratio is a specified but general function of the parameters. This development of assumption (1) is discussed by Anderson (1981b) where likelihood methods of estimation are given.

A major advantage of the penalized likelihood method is that smooth estimates of the unknown function are given, avoiding assumptions or approximations about its form. Instead, the roughness penalty function and the smoothing constant K have to be

selected. An advantage of this approach for inference is that it avoids the need to choose between conditional, partial and marginal likelihood. Judgement of the asymptotic efficiency of penalized likelihood must await the results of further basic work, but we believe that it provides a valuable methodological tool.

V. Blair had a Science Research Council Studentship.

Appendix

The spline estimate of $f(.)$

The maximization of the function in (13) can be achieved using calculus of variation techniques in a similar way to Silverman (1978). Suppose that x_i is the ith order statistic ($i = 1, \ldots, n$) and that $I = [x_0, x_{n+1}]$, where x_0 or x_{n+1} may be finite or infinite. For each strictly positive value of K the estimate $\tilde{f}(.)$ of $f(.)$ is a continuous function that has a piecewise continuous first derivative that is zero at the ends of the interval I and has a discontinuity of $-1/k$ at x_i ($i = 1, \ldots, n$). Further, it can be shown that $\tilde{f}(.)$ has the form

$$\tilde{f}(x) = (A_i e^{\mu x} + B_i e^{-\mu x})^2 \quad (x_i < x < x_{i+1};\ i = 0, \ldots, n),$$

where for convenience, we take $\mu > 0$. The constants μ, $\{A_i\}$ and $\{B_i\}$ are given by the following equations holding for $i = 0, \ldots, n-1$:

$$A_{i+1} e^{\mu x_{i+1}} + B_{i+1} e^{-\mu x_{i+1}} = A_i e^{\mu x_{i+1}} + B_i e^{-\mu x_{i+1}},$$

$$A_{i+1} e^{\mu x_{i+1}} - B_{i+1} e^{-\mu x_{i+1}} = A_i e^{\mu x_{i+1}} - B_i e^{-\mu x_{i+1}} - 1/\{k\mu(A_i e^{\mu x_{i+1}} + B_i e^{-\mu x_{i+1}})\};$$

$$A_0 e^{\mu x_0} - B_0 e^{-\mu x_0} = 0 \quad (x_0 > -\infty) \quad \text{or} \quad B_0 = 0 \quad (x_0 = -\infty);$$

$$A_n e^{\mu x_{n+1}} - B_n e^{-\mu x_{n+1}} = 0 \quad (x_{n+1} < \infty) \quad \text{or} \quad A_n = 0 \quad (x_{n+1} = \infty).$$

The solution to this system of equations can be found using a fairly straightforward computer algorithm.

References

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

Anderson, J. A. (1979). Compound logistic distributions. *Biometrika* **66**, 17–26.

Anderson, J. A. (1981a). Logistic discrimination. In *Handbook of Statistics*, **2**, Ed. P. R. Krishnaiah. New York: North-Holland.

Anderson, J. A. (1981b). Robust inference using logistic models. *Bull. Int. Statist. Inst.* **48**. *To appear.*

Anderson, J. A. & Senthilselvan, A. (1980). Smooth estimates for the hazard function. *J. R. Statist. Soc.* B **42**, 322–7.

Anderson, J. A., Whaley, K., Williamson, J. & Buchanan, W. W. (1972). A statistical aid to the diagnosis of *kerato-conjunctivitis sicca*. *Quart. J. Med.* **41**, 175–89.

Breslow, N. E. & Powers, W. (1978). Are there two logistic regressions for retrospective studies? *Biometrics* **34**, 100–5.

Clayton, J. K., Anderson, J. A. & McNicol, G. P. (1976). Pre-operative prediction of post-operative deep vein thrombosis. *Br. Med. J.* **2**, 910–2.

Cox, D. R. (1970). *The Analysis of Binary Data*. London: Methuen.

Cox, D. R. (1972). Regression models and life tables (with discussion). *J. R. Statist. Soc.* B **32**, 283–301.

Cox, D. R. (1975). Partial likelihood. *Biometrika* **62**, 269–76.

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

Good, I. J. & Gaskins, R. A. (1971). Non-parametric roughness penalties for probability densities. *Biometrika* **58**, 255–77.

Kalbfleisch, J. D. & Prentice, R. L. (1973). Marginal likelihoods based on Cox's regression and life model. *Biometrika* **60**, 267–78.

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

RICHARDS, F. S. G. (1961). A method of maximum-likelihood estimation. *J. R. Statist. Soc.* B **23**, 469–75.

SILVERMAN, B. W. (1978). Density ratios, empirical likelihood and cot death. *Appl. Statist.* **27**, 26–33.

WAHBA, G. & WOLD, S. (1975). A completely automatic french curve: fitting spline functions by cross validation. *Comm. Statist.* **4**, 1–17.

[*Received February* 1980. *Revised August* 1981]