Weighted likelihood estimating equations: The discrete case with applications to logistic regression

Marianthi Markatou[a,*], Ayanedranath Basu[b], Bruce Lindsay[c]

[a] *Department of Statistics, Columbia University, New York, NY 10027, USA*

[b] *Department of Mathematics, University of Texas, Austin, TX 78712, USA*

[c] *Department of Statistics, The Pennsylvania State University, University Park, PA 16802, USA*

Received 1 October 1994; revised 1 May 1995

Abstract

We discuss a method of weighting the likelihood equations with the aim of obtaining fully efficient and robust estimators. We discuss the case of discrete probability models using several weighting functions. If the weight functions generate increasing residual adjustment functions then the method provides a link between the maximum likelihood score equations and minimum disparity estimation, as well as a set of diagnostic weights and a goodness of fit criterion. However, when the weights do not generate increasing residual adjustment functions a selection criterion is needed to obtain the robust root.

The weight functions discussed in this paper do not automatically downweight a proportion of the data; an observation is significantly downweighted only if it is inconsistent with the assumed model. At the true model, therefore, the proposed estimating equations behave like the ordinary likelihood equations. We apply our results to several discrete models; in addition, a toxicology experiment illustrates the method in the context of logistic regression.

AMS classification: 62J12; 62A10; 62G35

Keywords: Generalized linear models; Likelihood approach; Residual adjustment functions; Robustness

1. Introduction

Robustness of estimation procedures and hypotheses tests has been a major concern in recent statistical literature. Huber (1964, 1973) demonstrated the existence of M-estimators of location and regression with minimax asymptotic variance over a specified neighborhood of a given distribution shape. This generated the minimax approach to robustness. Hampel (1968, 1974) assessed the robustness of an estimator by viewing it as a functional and examining the behavior of the first Gateaux derivative of this functional at the ideal model distribution. This generated the important concept of influence function and the infinitesimal approach to robustness. Following the above two schools of robustness a number of important papers have been written extending the ideas from the location-scale case to the regression case.

The problems that have extensively been treated in the literature are characterized by the presence of the invariance structure. Linear models, and generally problems in which ordinary least squares is the basic estimation technique, have received considerable attention. However, less attention has been paid to problems that do not have invariance structure. These are precisely the problems which motivated the current work. For this type of problems, such as exponential family models and mixture models, ordinary least squares does not constitute the basic fitting technique and standard robustness methods are either not available or are not well developed.

The approach we take is based on the idea of modifying the usual likelihood equations to achieve efficient estimates with good breakdown properties. To do this, we replace the maximum likelihood score equations with *weighted score equations*, in which the weights are functions of appropriately defined residuals. The role of the weights is to reduce the impact of outlying observations on the score equations.

The description of resistant alternatives to maximum likelihood estimation at least for generalized linear models is not new; see Besag (1981) and Pregibon (1982). A Royal Statistical Society, Series B, discussion paper by Green (1984), while extensively discussing the theory and use of iteratively reweighted least squares for maximum likelihood, suggested the replacement of the usual maximum likelihood score equations with weighted score equations. The weight functions proposed by Green (1984) are functions of the deviance $\Delta(x;\beta) = -2\{\ln f(x;\beta) - \sup_\tau \ln f(x;\tau)\}$. Lenth and Green (1987) use weights $w\{\Delta^{1/2}(x;\beta)\} = \psi_c(\Delta^{1/2}(x;\beta))/\Delta^{1/2}(x;\beta)$ where ψ_c is the Huber's psi-function. Field and Smith (1994) use weights based on the empirical cumulative distribution function.

In the present paper, we discuss a new method to construct weighted likelihood score equations and study the corresponding estimators in discrete models. The case of continuous distributions is treated in a sequel paper by Markatou et al. (1995). We note here that M-estimators for discrete distributions were proposed by Simpson et al. (1987), while minimum Hellinger distance estimators as well as other robust estimators which are fully efficient were studied by Beran (1977a, b, 1982) and Simpson (1987). Simpson (1989) also proposed the robust Hellinger deviance test. Section 2 of the present paper develops the methodology and discusses the properties of the estimators in discrete models. Section 3 applies the techniques to a few specific models and also illustrates the methodology in the context of a toxicology experiment with the logistic regression model.

2. The weighted likelihood equations

2.1. Methodology

Suppose that $\{X_1, X_2, \ldots, X_n\}$ is a random sample with probability mass function $m_\beta(x)$ defined on the sample space $R_X = \{0, 1, \ldots, T\}$ where T is possibly infinity. Let $u(x; \beta) = \nabla_\beta \ln(m_\beta(x))$ be the maximum likelihood score function, ∇_β being the gradient with respect to β. Assuming that the family of distributions is regular, the maximum likelihood estimator for β is a solution of the set of equations $\sum_{i=1}^{n} u(X_i; \beta) = 0$. Given any point x in the sample space we construct a weight depending on x, the assumed probability model M_β, and the empirical cumulative distribution function $\hat{F}$, say $w(x; M_\beta, \hat{F})$. We will consider solutions to the *weighted likelihood equations*

$$\sum_i w(X_i; M_\beta, \hat{F}) u(X_i; \beta) = 0. \tag{2.1}$$

The weight function $w(x; M_\beta, \hat{F})$ is selected such that it has a value close to 1 if there is no evidence of model violation at x from the empirical distribution function. It has a value close to 0 or exactly 0 at X_i if the empirical cumulative distribution function indicates lack of fit at or near X_i, $i = 1, 2, \ldots, n$. Thus, the role of the weight function is to downweight points that are inconsistent with the assumed model. However, to be efficient, the method should not automatically downweight a proportion of the data; if the assumed model is correct, the weight assigned to each observation should asymptotically be equal to 1.

To describe an observation as an outlier we need to define an appropriate set of residuals. The proper definition of residuals depends on the purpose for which they are intended. If there is a structural relationship between the observations and the parameters, such as in the linear models, then it is appropriate to expect a one-to-one correspondence between residuals and observations. In our context, the relationship between parameters and observations is probabilistic. Hence, instead of using a geometric interpretation, in which a point is an outlier if it is far away from the bulk of the data, we will call an observation an outlier if the offending value would be very unlikely to occur if the fitted model were true. Such a probabilistic outlier is called a *surprising* observation (Lindsay 1994), and occurs in locations t with small probabilities $m_\beta(t)$ under the model. We note that Davies and Gather (1993) also defined outliers in terms of their position relative to the model that most of the observations follow.

For any value t in R_X, define the *Pearson residual* δ as

$$\delta(t) = \frac{d(t)}{m_\beta(t)} - 1, \tag{2.2}$$

where $d(t)$ is the proportion of sample observations with value t and $m_\beta(t)$ is the corresponding probability under the hypothesized model. Note that

$$P^2 = n\sum_t m_\beta(t)\delta^2(t) \tag{2.3}$$

is the Pearson's chi-squared statistic for the goodness of fit of the model $m_\beta(t)$. The Pearson residual takes values in the interval $[-1,\infty)$. If the observed proportion of values at t is the same as the probability of observing t under the assumed model, then $\delta(t) = 0$; when the model is correctly specified, $\delta(t)$ converges to 0 almost surely. Also $\delta(t) = -1$ if no data is observed at t.

Our method downweights observations that have large Pearson residuals. Large Pearson residuals correspond to surprising observations. This downweighting scheme is distinctly different from downweighting observations that have large influence on the maximum likelihood estimator, in the sense that their presence or absence substantially changes the value of the estimator. Influential observations in the latter sense correspond to large values of the score function.

The weights we use are functions of the Pearson residuals and they are defined as

$$w(t, M_{\hat\beta}, \hat F) = w(\delta(t)) = \frac{A(\delta(t))+1}{\delta(t)+1}, \tag{2.4}$$

where A is a strictly increasing, twice differentiable function defined on $[-1,\infty)$, with the properties $A(0)=0$, $A'(0)=1$. In particular we will choose $A(\cdot)$ to be a *residual adjustment function* (RAF) (Lindsay, 1994) which arises naturally in some density-based minimum distance methods. Lindsay has shown that the choice of an RAF may have a dramatic effect on the robustness of the corresponding estimators. Roughly speaking, the effect of the RAF on Pearson residuals is similar to the effect of ψ_c function (in M-estimation) on the ordinary least-squares residuals. In the next subsection we will describe the connection of the weighted likelihood estimators with the minimum disparity estimators of Lindsay (1994). This will explain the rationale behind the choice of the weights (2.4), and will help us to describe the robustness properties of the new estimators using previously established results.

2.2. *Minimum disparity estimation*

In this subsection we review some fundamental background results. An extensive discussion of this topic can be found in Lindsay (1994).

Minimum disparity estimators of the parameter β are obtained by minimizing a disparity, which is a density-based "distance" between the empirical density $d(\cdot)$ and the underlying density $m_\beta(\cdot)$. Such a measure is defined by

$$\rho_G(d, m_\beta) = \sum_t G(\delta(t))m_\beta(t), \tag{2.5}$$

where G is a real-valued, thrice differentiable strictly convex function on $[-1,\infty)$ with $G(0)=0$. The function $G(\delta) = (\delta+1)\ln(\delta+1)$ generates a Kullback–Leibler divergence that is minimized by the maximum likelihood estimator. The function $G(\delta) =$

$[(\delta+1)^{1/2}-1]^2$ corresponds to the squared Hellinger distance. Under differentiability and appropriate regularity conditions, the minimum disparity estimating equations have the form

$$\sum_t A(\delta(t))\nabla_\beta m_\beta(t) = 0, \tag{2.6}$$

where $A(\delta) = (1-\delta)G'(\delta) - G(\delta)$, G' being the derivative of G. Without changing the estimating properties of the disparity, the function A can be recentered and scaled so that $A(0) = 0$ and $A'(0) = 1$. This standardized version of the function A is called the RAF of the disparity ρ_G and controls the theoretical properties of the minimum disparity estimators. The linear RAF, $A(\delta) = \delta$, corresponds to maximum likelihood, while $A(\delta) = 2\{(\delta+1)^{1/2} - 1\}$ corresponds to the Hellinger distance.

Since $A(\delta) = \delta$ corresponds to maximum likelihood, it is clear that the degree of robustness of the minimum disparity estimators (relative to maximum likelihood) will depend on how much $A(\delta)$ deviates from the linear RAF. If $A(\delta)$ is much smaller than δ when δ is large, then large δ outliers will have small impact on the parameter estimates.

Since $\sum_{t=1}^{T} \nabla_\beta m_\beta(t) = 0$, and $u(t,\beta) = \nabla_\beta m_\beta(t)/m_\beta(t)$, using (2.2) one can rewrite the estimating equation (2.6) as

$$\sum_{t=1}^{T} \frac{A(\delta(t)) - 1}{\delta(t)+1}\, u(t,\beta)d(t) = 0. \tag{2.7}$$

However, using $w(x,M_\beta,\hat{F}) = w(\delta(x)) = [A(\delta(x)) - 1]/[\delta(x)+1]$ (see Eq. (2.4)), and by rewriting the sum in t in terms of the sum in i, we see that the estimating equation (2.6) is exactly equivalent to the estimating equation in (2.1). Thus defining the weights as in (2.4) guarantees that the weighted likelihood estimator is a root of the minimum disparity estimating equation (2.6). The solution of the weighted likelihood equation can be obtained iteratively by calculating new weights at each stage and solving the estimating equation treating the weights as fixed constants.

Lindsay (1994) has shown that the breakdown properties of the estimators are determined by the tails of $A(\delta)$. In a model with finite Fisher's information, any estimator with $A(\delta) \sim \delta^{1/2}$ as δ becomes infinite has a 50% breakdown point. Also, the curvature parameter $A_2 = A''(0)$ can be shown to be a measure of the trade-off between efficiency and robustness in a second-order asymptotic sense – large negative values of A_2 lead to higher robustness. In a weighted likelihood context, definition (2.4) and the choice of the function $A(\delta)$ guarantee that the weights are all converging to 1 at the model, indicating the asymptotic efficiency of the weighted likelihood estimator.

On the other hand, the RAF of a disparity like the Hellinger distance, for which $A(\delta) \ll \delta$ for large positive δ, the weight function can severely downweight large Pearson residuals. For the sake of simplicity and interpretation, however, we can truncate the weights to constrain them in the closed interval $[0,1]$. This can be done by using $w(x,M_\beta,\hat{F}) = \min\{[A(\delta(x)) + 1]^+/(\delta(x)+1), 1\}$. Asymptotically this makes no difference in the estimation procedure under the model. From the form of the weights it is clear that one can recover the RAF given the weights as

$$A(\bar{\delta}) = -1 + (\bar{\delta} + 1)w(\bar{\delta}). \tag{2.8}$$

Lindsay (1994, Eq. (15)) has shown that as long as the function $A(\delta)$ is increasing, the associated estimating equation (2.6) corresponds to a minimum disparity problem. Since the truncation $w(x, M_\beta, \hat{F}) = \min\{[A(\delta(x)) + 1]^+/(\delta(x) - 1), 1\}$ preserves the increasing nature of $A(\bar{\delta})$ in (2.8), the corresponding procedure still generates fully efficient estimates.

2.3. Robustness properties

The robustness of a statistic can be studied in terms of its influence function and its breakdown point. The concept of influence function was introduced by Hampel (1968, 1974) and describes the bias caused by infinitesimal contamination on the value of the statistic. If $T(F)$ is a functional corresponding to an estimator, the influence function of $T(F)$ is given by $\mathrm{IF}(x; T, F) = \lim[T((1-\varepsilon)F + \varepsilon\Delta_x) - T(F)]/\varepsilon$, where the limit is taken as ε approaches 0 from above and Δ_x is the distribution that concentrates all its mass at the point x.

The breakdown point characterizes the global stability of an estimator. There are both asymptotic (Hampel, 1971; Huber, 1981) and finite-sample (Hodges, 1967; Donoho and Huber, 1983) versions of the concept of breakdown. Roughly speaking, the breakdown point of an estimator is the distance from the assumed distribution of the data beyond which the estimator becomes totally uninformative.

In our case the functional $\beta_w(F)$ corresponding to the weighted likelihood estimator will be a chosen element of the solution set of the equation

$$\int w(x; M_\beta, F)u(x; \beta)\,\mathrm{d}F(x) = 0. \tag{2.9}$$

Note that if $F = M_{\beta_0}$, then the true value of β, β_0, is among the solutions of equation (2.7). Therefore, the method is Fisher consistent for β if the root is chosen appropriately. We will discuss the root selection in the examples section.

To determine the influence function set $F_\varepsilon(x) = (1 - \varepsilon)F(x) + \varepsilon\Delta_y(x)$, $0 < \varepsilon < 1$. Let f_ε and f be the densities corresponding to F_ε and F, and the Pearson residuals are $\delta(t) = f(t)/m_{\beta_0}(t) - 1$, $\delta_\varepsilon(t) = f_\varepsilon(t)/m_{\beta_\varepsilon}(t) - 1$, where $\beta_0 = \beta_w(F)$ and $\beta_\varepsilon = \beta_w(F_\varepsilon)$.

Proposition 1. *The influence function of the weighted likelihood estimator is given by*

$$\frac{\partial}{\partial\varepsilon}\beta_\varepsilon|_{\varepsilon=0} = \beta'_w(y) = A(F)B(y; F),$$

where

$$A(F) = \left[\int w'(\delta(t))u(t; \beta_0)u^{\mathrm{T}}(t; \beta_0)(\delta(t) + 1)\,\mathrm{d}F(t) + \int w(\delta(t))(-\nabla u(t; \beta_0))\,\mathrm{d}F(t)\right]^{-1},$$

$$B(y;F) = w(\delta(y))u(y;\beta_0) + w'(\delta(y))u(y;\beta_0)(\delta(y)+1)$$
$$- \int w'(\delta(t))f(t)\frac{u(t;\beta_0)}{m_{\beta_0}(t)}\,\mathrm{d}F(t).$$

At the model $F = M_{\beta_0}$ this reduces to the simple form

$$\beta'_w(y) = \left[\int -\nabla_\beta u(t;\beta_0)\,\mathrm{d}M_{\beta_0}(t)\right]^{-1} u(y;\beta_0),$$

which is exactly the same as the influence function of the maximum likelihood estimator.

(In future references, we will denote the corresponding second derivative evaluated at $\varepsilon = 0$ by $\beta''_w(F)$.)

Proof. Implicit differentiation of $\int w(\delta_\varepsilon(t))u(t;\beta_\varepsilon)\,\mathrm{d}F_\varepsilon(t) = 0$ leads to the equation

$$\int w'(\delta_\varepsilon(t))\delta'_\varepsilon(t)u(t;\beta_\varepsilon)\,\mathrm{d}F_\varepsilon(t) + \int w(\delta_\varepsilon(t))\nabla u(t;\beta_\varepsilon)\,\mathrm{d}F_\varepsilon(t)\cdot\beta'_\varepsilon$$
$$+ \int w(\delta_\varepsilon(t))u(t;\beta_\varepsilon)\,\mathrm{d}(\Delta_y - F) = 0,$$

where

$$\delta'_\varepsilon(t)|_{\varepsilon=0} = \frac{\partial}{\partial\varepsilon}\{f_\varepsilon(t)/m_{\beta_\varepsilon}(t) - 1\}|_{\varepsilon=0}$$
$$= [I(t = y) - f(t) - u^{\mathrm{T}}(t;\beta_0)\cdot\beta'\cdot f(t)]/m_{\beta_0}(t).$$

Thus the influence function has the form $\beta' = \beta'(y) = A(F)B(y;F)$, where $A(F)$ and $B(y;F)$ are as above. If $F = M_{\beta_0}$, then $\delta = 0$, $w(\delta) = 1$, $w'(\delta) = 0$, and the influence function is exactly the same as that of maximum likelihood.

The above analysis indicates that the method provides an efficient estimator of the model parameters when the model is true. However, since the influence function of the weighted likelihood estimators is identical to that of the maximum likelihood estimator at the model, it can potentially be unbounded. In fact, all the weighted likelihood estimators are exactly equivalent to the maximum likelihood estimator at the model up to the first order of analysis. Any distinction between these methods, therefore, has to be made through a higher-order analysis. Later in this section we show that a second-order analysis of the bias function demonstrates the limitations of the first-order analysis based on the influence function and illustrates why the influence function can be a misleading indicator of the robustness properties of the weighted likelihood estimators.

While it cannot be employed to assess the robustness of the new estimators, the influence function is useful for obtaining their standard errors. From the previous calculation, the asymptotic variance of $n^{1/2}$ times the estimators is

$$\Sigma_\beta = A(F)E\{B(Y;F)B(Y;F)^{\mathrm{T}}\}A^{\mathrm{T}}(F), \tag{2.10}$$

which can be estimated consistently in the "sandwich" fashion by

$$\hat{\Sigma}_\beta = A(\hat{F})\left[\frac{1}{n}\sum B(X_i;\hat{F})B(X_i;\hat{F})^{\mathrm{T}}\right]A^{\mathrm{T}}(\hat{F}). \tag{2.11}$$

Viewed as a function of ε, $\Delta\beta(\varepsilon) = \beta(F_\varepsilon) - \beta(F)$ represents the bias of the functional $\beta(F)$ under contamination. We now show that first-order approximation through the influence function may give an unreliable prediction of the bias for the new estimators. By a simple Taylor series expansion, the ratio of the quadratic-to-linear approximation of $\Delta\beta(\varepsilon)$ can be seen to be $1+[\beta''(y)/\beta'(y)](\frac{1}{2}\varepsilon)$. Therefore, if the amount of contamination ε is greater than $\varepsilon_{\text{crit}} = |\beta'(y)/\beta''(y)|$, and $\beta'(y)$ and $\beta''(y)$ are opposite in sign, the two approximations will differ by more than 50%, with the quadratic approximation predicting lower bias.

Consider a scalar parameter β for the model $\{m_\beta\}$, and let $i(\beta)$ be the Fisher information. From Proposition 1, $\beta'(y) = [i(\beta)]^{-1}u(y;\beta)$; a straightforward calculation involving the second derivative of the estimating function under contamination gives

$$\beta''(y) = \beta'(y)[i(\beta)]^{-1}[f_1(y) + A_2 f_2(y)], \tag{2.12}$$

where

$$\begin{aligned} f_1(y) &= 2\nabla u(y;\beta) - 2E[\nabla u(x;\beta)] + \beta'(y)E[\nabla^2 u(x;\beta)], \\ f_2(y) &= \frac{i(\beta)}{m_\beta(y)} - 2u^2(y;\beta) + E[u^3(x;\beta)]\frac{u(y;\beta)}{i(\beta)}. \end{aligned} \tag{2.13}$$

Moreover, if the model is a one-parameter exponential family, with β being the mean value parameter, $f_1(y) = 0$ and hence $\varepsilon_{\text{crit}} = |A_2 Q(y)|^{-1}$ with

$$Q(y) = \frac{1}{m_\beta(y)} + \frac{E[(X-\beta)^3](y-\beta)}{\{E[(X-\beta)^2]\}^2} - 2\frac{(y-\beta)^2}{E[(X-\beta)^2]}. \tag{2.14}$$

If we use the approximation $Q(y) \approx 1/m_\beta(y)$, we get $\varepsilon_{\text{crit}} \approx m_\beta(y)/|A_2|$. If $A_2 < 0$, then the signs of $\beta'(y)$ and $\beta''(y)$ are opposite; in this case, whenever $\varepsilon > \varepsilon_{\text{crit}}$, the quadratic approximation will predict a bias which is smaller by 50% or more compared to the bias predicted by the first-order influence function approach, showing the limitation of the latter in this case.

2.4. The weights: goodness of fit tests and calibration

One of the important issues in the development of the theory of weighted likelihood equations is the calibration of the weights. That is, how do we select an appropriate weight function, and measure its impact on our procedure. Some key insights into this question can be obtained by examining the variability of the weights when the model is correct. Let $\bar{w} = n^{-1}\sum w(X_i; M_{\hat{\beta}}, \hat{F})$. The following surprising result indicates that the final sum of fitted weights, in the multinomial case, is a chi-square goodness of fit test for the model, against the general multinomial alternative. Let $W_2 = -A''(0)$, and assume that $W_2 \neq 0$.

Proposition 2. *Under regularity conditions, in the k-cell multinomial model, when the model is correctly specified*

$$2n(1-\bar{w})W_2^{-1} - P^2 \to 0$$

in probability, where P^2 is the Pearson's chi-squared statistic.

Proof. Do a Taylor expansion of $w(\delta(t))$ about $\delta(t) = 0$ and obtain

$$n(1-\bar{w}) = \frac{W_2}{2}\left[n\sum_t d(t)\delta^2(t)\right] + o_p(1),$$

where the summed term is asymptotically equivalent to the Pearson chi-squared goodness of fit test, as the parameter estimate for β is asymptotically equivalent to the maximum likelihood estimator.

As a corollary, we note that the difference in weights between two competing nested models can be used as a chi-squared test of the smaller model against the larger.

The relevance of this result to the estimation process is that we can expect the sum of weights, when the model is correct, to be roughly equal in magnitude to $(n - n^*)$, with $n^* = \frac{1}{2}W_2^{-1}(k - 1 - \dim(\beta))$. Thus n^* reflects, at least in an intuitive sense, the loss of sample size necessary to achieve the improved robustness properties.

In general, a higher degree of robustness may be achieved by making the weights sharper; this can be accomplished by choosing a higher power of the weights. To include such cases within the family of weighted likelihood estimators, we define the general weighted likelihood estimation method through the estimating equation

$$\sum_{i=1}^{n}[w(X_i, M_{\hat{\beta}}, \hat{F})]^k u(X_i, \beta) = 0, \quad k \geqslant 0. \tag{2.15}$$

We recover the maximum likelihood score equations when $k = 0$, get the weighted likelihood equations in (2.1) for $k = 1$, and for $k > 1$ get estimating equations which provide stronger downweighting of the outlying (high Pearson residual) cells. The estimating equation can again be iteratively solved by creating new weights at each stage. Note that at the model the weights are still converging to 1 for any finite k. It is easy to generalize the influence function analysis of Proposition 1 and show, by taking a derivative of the estimating equation (2.15) under the contamination F_ε, $F = M_{\beta_0}$, that the resulting estimator still has the same influence function as the maximum likelihood estimator at the model. However, the larger the value of k, the greater is the amount of downweighting for an outlying observation. A straightforward analysis similar to the second-order bias function analysis of Section 2.3 shows that the second-order approximation to the bias has a similar form given by Eqs. (2.12) and (2.13), but now the term $f_2(y)$ is k times the term obtained in (2.13). Therefore, for the mean-value parameter of the one-parameter exponential family model, $\varepsilon_{\text{crit}}$ is now $|kA_2Q(y)|^{-1}$, where $Q(y)$ is defined in (2.14). Thus $\varepsilon_{\text{crit}}$ is a decreasing function of k, and larger values of k will predict smaller bias. This second-order bias analysis parallels that of Lindsay (1994, Section 4).

Rewriting Eq. (2.15) in the form (2.6), one can see that this corresponds to an RAF $A(\delta) = -1 + (\delta + 1)[w(\delta)]^k$; this manipulation of the weights provides no theoretical problems as long as this corresponds to an increasing $A(\delta)$, in which case it follows from Lindsay (1994, Eq. (15)) that there exists a convex $G(\delta)$ that generates a disparity measure as in (2.5), with corresponding estimating equation given by (2.15). In particular when starting from Eq. (2.15), we get $A'(\delta) = (1 + \delta)kw(\delta)w'(\delta) + [w(\delta)]^k$, and following Lindsay can regenerate the function

$$G(\delta) = \int_0^{\delta} ([w(t)]^k - 1)\,\mathrm{d}t + \int_0^{\delta} \int_0^{s} \frac{[w(t)]^k}{(1-t)}\,\mathrm{d}t.$$

Since when the weights produce an increasing RAF there is a one-to-one correspondence between the set of likelihood equations and a minimization problem the results of Lindsay (1994) can be used to show that the weighted likelihood estimators have high breakdown points; in particular when $A(\delta) \sim \delta^{1/2}$ in the weight function, the breakdown point of the estimators can be 50%. For comparison, Christmann (1994) showed that in logistic regression with large strata a modification of Rousseeuw's least median of squares estimator has a finite sample breakdown point of approximately $\frac{1}{2}$.

In case the function $A(\delta)$ is not an increasing function, as it would generally be the case for arbitrary power of the weight k, the above procedure generates a criterion function via (2.5). Although it *no longer is a formal disparity measure*, it asymptotically has a local minimum at the true parameter. To select the robust root the criterion function generated via (2.5) is used.

While we are still using minimum disparity ideas to select the robust root in the case of multiple solutions to the estimating equations, there are several important differences between minimum disparity estimation and weighted likelihood estimation: (a) The method establishes a link between the maximum likelihood score equation and the minimum disparity estimation equations via the weights. (b) Apart from robust estimates, the method provides a set of diagnostic weights. For observations that are consistent with the model we expect the weights to be close to 1; if the data set contains observations absolutely inconsistent with the model their weights are close to 0, so that we expect solutions much like the maximum likelihood estimator of the subset of the data excluding the most unlikely observations. (c) The set of weights are extremely useful in testing goodness of fit and calibration.

3. Examples

In this section we apply the theory developed of some discrete models to demonstrate the robustness properties of the procedure. We also discuss the application of our methodology to examples from logistic regression. The goal of these numerical examples is to study the behavior of the roots of the set of estimating equations and to show how we can construct robust and efficient estimators. All the numerical calculations presented in this section were carried out on a SUN workstation, and the programs were written in Splus and Fortran.

When the true density is given by $0.5f(x) + 0.5g(x)$ where each of the component densities are in the neighborhood of the assumed model $\{m_\beta\}$, we measure robustness by the existence of a root at or near one of the components involved. When the contamination proportion is less than 0.5, we measure robustness by the existence of a root at or near the component with the larger mass. The examples presented here show that the robustness properties of the proposed methods are dependent on the power of the weight k and the separation of the two populations involved. Although the set of estimating equations may not have a unique solution, we can select the robust root by going back to the corresponding minimization problem. In particular our examples show the following:

(a) When the separation between the two components, as measured by the distance between their corresponding centers is sufficiently large, there always exists a robust root at or near the component with the larger mass.

(b) When the true density is a mixture of two densities with equal proportion and the separation between the two populations is small, a robust root may be obtained by increasing the power of the weight.

(c) The number of roots may be a function of the amount of contamination.

For the following calculations, we let $m_\beta(x)$ correspond to the Poisson(β) model. Initially, we replace the data by the mixture density $d(x) = (1-\varepsilon)m_2(x) + \varepsilon m_{15}(x)$, and attempt to determine the estimate of β using the weighted likelihood method, for the residual adjustment function $A(\delta) = 2\{(\delta+1)^{1/2} - 1\}$, which corresponds to the Hellinger distance.

Fig. 1 shows a plot of the weighted likelihood estimating functions against various values of the parameter β and various levels of contamination ε. The power of the weights used here is $k = 1$. Note that for $\varepsilon < 0.3$ there is only one single root near 2. For $\varepsilon > 0.3$ more roots appear, such that if $\varepsilon = 0.5$ there is one root in the neighborhood of 2, another root in the neighborhood of 15 and a third root in the neighborhood of the maximum likelihood estimator. The point of the graph is to demonstrate that the number of roots changes as the amount of contamination increases provided there is sufficient separation in the two populations; in this example there is always a robust root in the neighborhood of 2 if the proportion of contamination is less than of 0.5. If the two populations are not sufficiently separated, then this may require a higher power of the weights. The role of the power of the weights is clearly illustrated in Figs. 2 and 3.

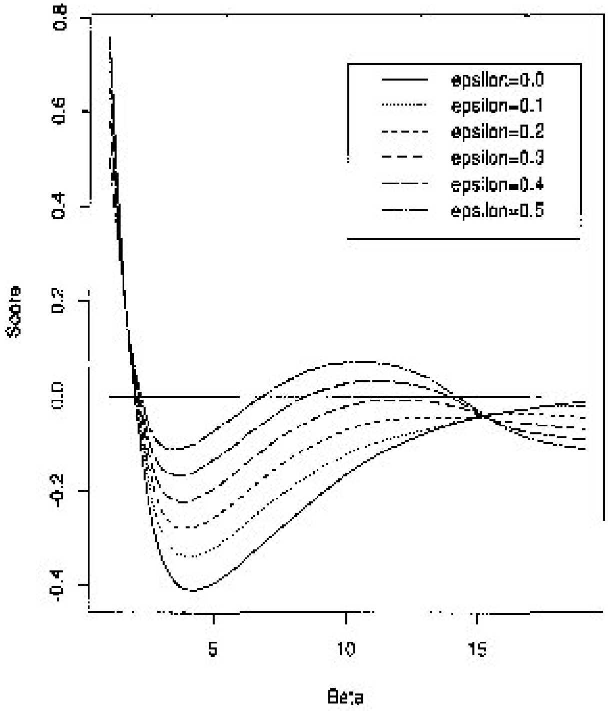

Fig. 1. The model is Poisson(β), and the data is the $(1-\varepsilon)$ Poisson(2) + ε Poisson(15) density. The weighted likelihood score functions are plotted against β, for several values of ε, the contamination proportion.

Fig. 2 is the graph of the weighted likelihood estimating functions against β, but this time $d(x)$ is the $0.5m_5(x) - 0.5m_{10}(x)$ mixture density. When the power of the weights is equal to 1, the set of the estimating equations has only one root, which is close to the MLE, 7.5. However, as one increases the power of the weight more roots appear. Thus, for $k = 11$ we again observe three roots, one close to 5, a second close to the MLE 7.5, and a third close to the mean of the Poisson(10) component. However, if the components have a larger separation, multiple roots appear for a smaller power of weights, as is seen in Fig. 3.

Fig. 3 shows the graph of the weighted likelihood estimating functions against β when we replace $d(x)$ with the $0.5m_3(x) + 0.5m_{10}(x)$ mixture density. Notice here that the separation of the two populations is greater than in Fig. 2 as indicated by the difference of their means. We see that powers of 1 and 2 still produce one root while if $k \geqslant 3$ we observe three roots, one in the neighborhood of 3, a second in the neighborhood of the MLE and the third one in the neighborhood of 10. To select an appropriate root when there exists multiple roots, we recommend choosing the root which provides the minimum in the corresponding minimization problem.

We will now discuss the application of our methodology to an example from logistic regression. Traditionally, logistic regression models have been fitted to data obtained under experimental conditions, for example bioassay applications. Currently, logistic

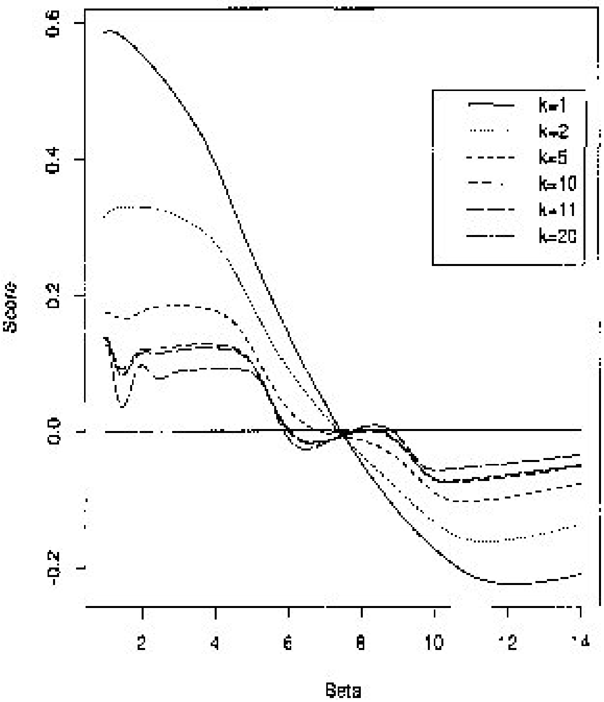

Fig. 2. The model is Poisson(β), and the data is the 0.5 Poisson(5) + 0.5 Poisson(10) density. The weighted likelihood score functions are plotted against β, for several values of k, the power of the weights.

regression models are being applied to data from observational studies. Data from such studies can be notoriously bad from the point of view of outlying y values as well as extreme values in the design space.

There is a considerable body of literature that discusses robustness issues associated with the logistic regression model. These are mostly adaptations of the usual linear regression model techniques. See, for example, Stefanski et al. (1986), Kunsch et al. (1989), Copas (1988), Carroll and Pederson (1993) and Christmann (1994) to mention a few. Here we discuss the application of our methodology to a particular situation of logistic regression.

Suppose we have I different covariate patterns and there are repeat observations at each covariate pattern x_i. Thus, there are n_i binary observations corresponding to x_i, and let Y_i be the number of observations equal to 1. We will denote the observed value of Y_i as y_i; hence Y_i has a binomial distribution with parameters (n_i, p_i). Let $d_i = y_i/n_i$ and let $p_i - \exp(\beta^{\mathrm{T}}x_i)/[1+\exp(\beta^{\mathrm{T}}x_i)]$. We may define the Pearson residual

$$\delta_1(x_i) = \frac{d_i}{\hat{p}_i} - 1, \tag{3.1}$$

where $\hat{p}_i = \exp(\hat{\beta}^{\mathrm{T}}x_i)/[1+\exp(\hat{\beta}^{\mathrm{T}}x_i)]$ is an estimate of p_i. We will like our estimation procedure to downweight the ith case if the corresponding Pearson residual $\delta_i(x_i)$ is

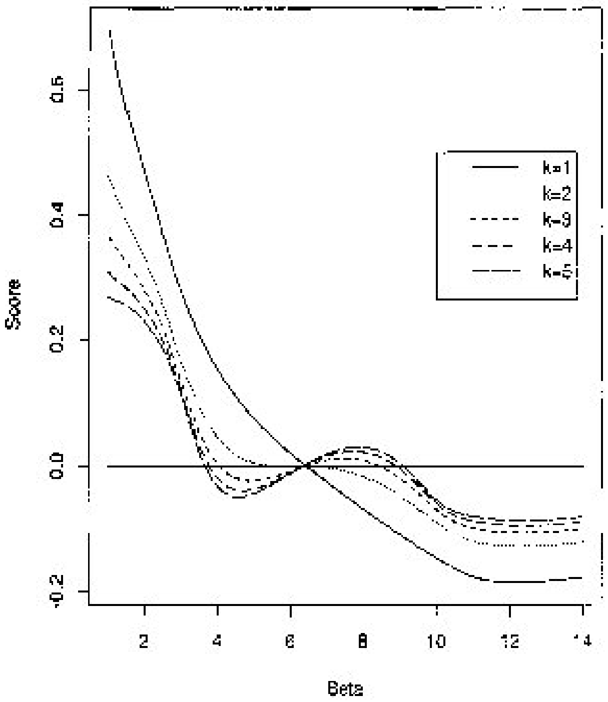

Fig. 3. The model is Poisson(β), and the data is the 0.5 Poisson(3) + 0.5 Poisson(10) density. The weighted likelihood score functions are plotted against β, for several values of k, the power of the weights.

large. However, consider the case where the observed value d_i is substantially smaller than the value predicted under the model. If we only downweight based on the value of the Pearson residual in (3.1), such a cell will generate a negative value of the Pearson residual δ_1, and may end up getting a weight close or equal to 1 for certain RAFs. Since any observation corresponding to a given covariate pattern x_i generates a two-cell Bernoulli distribution, these negative residuals for the "1-cells" (successes), actually may correspond to large positive residuals for the "0-cells" (failures). Hence, we also define

$$\delta_0(x_i) = \frac{1-d_i}{1-\hat{p}_i} - 1 \tag{3.2}$$

to be the residual associated with the "0-cell". Our estimators then are obtained by minimizing a weighted sum of distances. For a given convex function G we minimize, with respect to β,

$$\sum_{i=1}^{I} n_i \left[\frac{\exp(\beta^{\mathrm{T}} x_i)}{1+\exp(\beta^{\mathrm{T}} x_i)} G(\delta_1(x_i)) + \frac{1}{1+\exp(\beta^{\mathrm{T}} x_i)} G(\delta_0(x_i)) \right].$$

Since G is thrice differentiable the above optimization problem is equivalent to

$$\sum_{i=1}^{I} n_i [A(\delta_1(x_i)) - A(\delta_0(x_i))] \nabla_\beta \left\{ \frac{\exp(\beta^{\mathrm{T}} x_i)}{1-\exp(\beta^{\mathrm{T}} x_i)} \right\} = 0. \tag{3.3}$$

For illustration purposes note that the maximum likelihood estimating equations are

$$\sum_{i=1}^{I} n_i[\delta_1(x_i)\nabla_\beta p_i - \delta_0(x_i)\nabla_\beta(1-p_i)] = \sum_{i=1}^{I} n_i[\delta_1(x_i) - \delta_0(x_i)]\nabla_\beta p_i = 0. \qquad (3.4)$$

The equation (3.3) can be rewritten as

$$\sum_{i=1}^{I} [w_1(x_i)y_i(1-p_i) - w_0(x_i)(n_i - y_i)p_i]x_i = 0, \qquad (3.5)$$

where $w_1(x_i) = [A(\delta_1(x_i)) + 1]/(\delta_1(x_i) + 1)$ and $w_0(x_i) = [A(\delta_0(x_i)) + 1]/(\delta_0(x_i) + 1)$. This is because $\nabla_\beta p_i = p_i(1-p_i)x_i$. This can then be solved as a weighted likelihood: for current values of β and a given RAF we construct weights w_1, w_0 and perform a Newton–Raphson iteration to obtain a new value of $\hat{\beta}$. We repeat the procedure till convergence. Again the method provides fully efficient estimators of the parameter vector β.

We have seen that the role of the RAF is to shrink large residuals. It may also be of interest to study the behavior of the procedure for negative values of δ, particularly those near -1. To smoothly downweight such values (called Pearson inliers) we may use the *negative exponential* RAF (Lindsay 1994), defined as

$$A_{\mathrm{NE}}(\delta) = 2 - (2+\delta)e^{-\delta}. \qquad (3.6)$$

The negative exponential RAF downweights both positive and negative residuals relative to the maximum likelihood, in the sense that $|A(\delta)| \leqslant |\delta|$. We will apply both the negative exponential and Hellinger distance RAFs in the following example.

Example. This example involves data which resulted from a toxicological experiment conducted at the University of Waterloo, Canada, and are presented in O' Hara Hines and Carter (1993, p.13). Six different concentrations of the toxicant potassium cyanate (KSCN) were applied to 48 vials of trout fish eggs. Each vial contained between 61 and 179 eggs. The eggs in half the vials were allowed to water harden for several hours before the toxicant was applied (this is a process in which the surface of a fish egg becomes toughened after a few hours in water). For the remaining vials, the toxicant was applied immediately after fertilization. After 19 days of the start of the experiment the number of dead eggs in each vial was counted.

Treating the number of dead eggs in each vial as the response we fit a logistic regression model to the data with covariates for water hardening (0 if the toxicant was applied before water hardening and 1 if it was applied after), and for a linear and quadratic term in log-concentration of the toxicant. The quadratic term in log-concentration is used to describe a sharp increase in mortality caused by the two highest concentrations. The maximum likelihood estimators were used as the starting values.

Table 1 gives the weight of those cases that did not receive a weight of nearly or exactly 1. We have two columns of weights; column 1 corresponds to the weights of the response cells and column 0 to those of the nonresponse cells, for any given x_i.

Table 1
Weights for KSCN example

Case number	Negative exp. weights		Hellinger weights	
	Column 1	Column 0	Column 1	Column 0
12	0.8905032	1.0000000	0.8552530	0.9965012
13			0.0000000	0.9977647
14			0.3279548	0.9982219
28			0.4098746	0.9993424
32			0.0000000	0.9993965
34	0.8940925	1.0000000	0.8541044	0.9973958
35			0.2994592	0.9996005
36			0.7742827	0.9997611
37	0.8532178	1.0000000	0.8281611	0.9836277
38	0.6217503	1.0000000	0.7023431	0.8955089
39	0.7126061	1.0000000	0.7485037	0.9477368
40			0.0000000	0.9973932
41			0.8588582	0.9943002
42			0.3038337	0.9886016
43			0.0000000	0.9866767
44			0.0000000	0.9828289

The parameters α, β_1, β_2 and β_3 are the intercept, and the slope parameters associated with water-hardening, log-concentration and squared log-concentration.

The negative exponential RAF downweights observations 12, 34, 37, 38, and 39. After inspection of the data we see that observation 12 with weight 0.8905 and observation 34 with weight 0.8940 have, respectively, the highest number of dead eggs at concentration level 360, after and before water-hardening was applied. Observations 37, 38, and 39 at concentration level 720, prior to water-hardening, are also downweighted as having high mortality. Notice that observations 38 and 39 received the lowest weight. Examination of these observations showed that the mortality was high compared to all four replicates at the next higher concentration level at the same water-hardness level. O' Hara Hines and Carter (1993) pinpoint observations 38, 39 and 26 as possible outliers. Our method gives observation 26 a weight of nearly 1 indicating that it is consistent with the fitted model. An analogue of Cook's statistic also pinpointed observations 38, 39 as potential outliers.

When the Hellinger RAF is used for the construction of the weights, observations 13, 32, 40, 43 and 44 received a weight of 0. Examination of those observations reveals that observation 32 has a 0 response, while observations 40, 43 and 44 have the lowest mortality at concentration levels 720 and 1440 respectively at the same water-hardening level. For similar reasons observation 42 receives a weight of 0.3038, while observation 41 receives a weight of 0.8588. Observation 13, as having the lowest number of dead eggs at concentration level 720 and after water-hardening is applied, receives a weight of 0, indicating inconsistency with the fitted model.

Table 2
New effective sample size for the downweighted cases for NED RAF

Case number	New	Data	New effective sample size	Old sample size
	1-Cell	0-Cell		
12	15	86	101	103
34	19	124	143	145
37	25	70	95	99
38	33	56	89	109
39	29	59	88	99

Some comments about the interpretation of the final weights are appropriate here. Some insight into the role of the weights can be obtained by examining the solution as if the weights were fixed at final values — this gives the "data" for which the weighted likelihood estimator is the MLE. In our logistic regression problem, let y_i and $(n_i - y_i)$ represent the number of successes and the number of failures, respectively, at the ith case, and let w_{1i} and w_{0i} represent the final weights for the "1-cell" and the "0-cell" corresponding to the ith case; this generates the new "data" $(w_{1i}y_i, w_{0i}(n_i - y_i))$. The following are of note: (a) this decreases the effective sample size from n_i to the new effective sample size $(w_{1i}y_i + w_{0i}(n_i - y_i))$, and (b) shifts the sample proportion d_i to $(w_{1i}y_i + w_{0i}(n_i - y_i))^{-1}w_{1i}y_i$, which reflects the extent to which the model disagrees with the observed proportion. Note that if $y_i = 0$ then w_{1i} is irrelevant, hence the role of the weights is hard to separate from cell counts.

The case of binary data with no replications at each x_i is currently being investigated by the authors.

Acknowledgements

The work of Marianthi Markatou was partially supported by an RGK CU01583501 Foundation Grant. She would like to thank Professor S.H. Lo for his helpful comments and suggestions. The work of A. Basu was partially supported by a Mathematics Department research award, University of Texas at Austin. The work of Bruce Lindsay was partially supported by NSF Grant DMS-9403847.

References

Basu, A., and B.G. Lindsay (1994). Minimum disparity estimation for continuous models: efficiency, distributions and robustness. *Ann. Inst. Statist. Math.* **46**, 683–705.

Beran, R. (1977a). Robust location estimates. *Ann. Statist.* **5**, 431–444.

Beran, R. (1977b). Minimum Hellinger distance for parametric models. *Ann. Statist.* **5**, 445–463.

Beran, R. (1982). Robust estimation in models for independent nonidentically distributed data. *Ann. Statist.* **10**, 415–428.

Besag, J. (1981). On resistant techniques and statistical analysis. *Biometrika* **68**, 463–469.

Carroll, R.J. and S. Pederson (1993). On robustness in the logistic regression model. *J. Roy. Statist. Soc. Ser. B* **55**, 693–706

Christmann, A. (1994). Least median of weighted squares in logistic regression with large strata. *Biometrika* **81**, 413–417.

Copas, J.B. (1988). Binary regression models for contaminated data. *J. Roy. Statist. Soc. Ser. B* **50**, 225–265.

Davies, L. and U. Gather (1993). The identification of multiple outliers. *J. Amer. Statist. Assoc.* **88**, 782–792.

Donoho, D.L. and P.J. Huber (1983). The notion of breakdown point. In: P.J. Bickel et al. Eds., *Festschrift Forerich L. Lehmann.* Wadsworth, Belmont, CA. 157–184.

Field, C. and B. Smith (1994). Robust estimation: a weighted maximum likelihood approach. *Int. Statist. Rev.* **62**, 405–424.

Green, P.J. (1984). Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives. *J. Roy. Statist. Soc. Ser. B* **46**, 149–192.

Hampel, F.R. (1968). Contributions to the theory of robust estimation. Ph.D dissertation, Dept. of Statistics, Univ. of California, Berkeley (unpublished).

Hampel, F.R. (1971). A general qualitative definition of robustness. *Ann. Math. Statist.* **42**, 1887–1896.

Hampel, F.R. (1974). The influence curve and its role in robust estimation. *J. Amer. Statist. Assoc.* **69**, 383–393.

Hodges, J.L. Jr. (1967). Efficiency in normal samples and tolerance of extreme values for some estimators of location. *Proc. 5th Berkeley Symp. on Mathematical Statistics and Probability.* Vol. 1, 163–186.

Huber, P.J. (1964). Robust estimation of a location parameter. *Ann. Math. Statist.* **35**, 73–101.

Huber P.J. (1973). Robust regression: asymptotics, conjectures and Monte Carlo. *Ann. Statist.* **1**, 799–821.

Huber P.J. (1981). *Robust Statistics.* Wiley, New York.

Kunsch, H., L.A. Stefanski and R.J. Carroll (1989). Conditionally unbiased bounded influence estimation in general regression models, with applications to generalized linear models. *J. Amer. Statist. Assoc.* **84**, 460–466.

Lenth, R.V. and P.J. Green (1987). Consistency of deviance-based M-estimators. *J. Roy. Statist. Soc. Ser. B* **49**, 326–330.

Lindsay, B.G. (1994). Efficiency versus robustness: the case for minimum Hellinger distance and related methods. *Ann. Statist.* **22**, 1081–1114.

Markatou, M., A. Basu and B.G. Lindsay (1995). Weighted likelihood estimating equations: the continuous case. Tech. Rept., Dept. of Statistics, Columbia University, New York.

O'Hara Hines, R.J. and E.M. Carter (1993). Improved added variable and partial residual plots for the detection of influential observations in generalized linear models. *Appl. Statist.* **42**, 3–20.

Pregibon, D. (1982). Resistant fits for some commonly used logistic models with medical applications. *Biometrics* **38**, 485–498.

Simpson, D.G. (1987). Minimum Hellinger distance estimation for the analysis of count data. *J. Amer. Statist. Assoc.* **82**, 802–807.

Simpson, D.G. (1989). Hellinger deviance tests: efficiency, breakdown points and examples. *J. Amer. Statist. Assoc.* **84**, 107–113.

Simpson, D.G., R.J. Carroll and D. Ruppert (1987). M-estimation for discrete data: asymptotic distribution theory and implications. *Ann. Statist.* **15**, 657–669.

Stefanski, L.A., R.J. Carroll and D. Ruppert (1986). Optimally bounded score functions for generalized linear models with applications to logistic regression. *Biometrika* **73**, 413–424.