J. R. Statist. Soc. B (1984),
46, *No.* 2, *pp.* 149–192

Iteratively Reweighted Least Squares for Maximum Likelihood Estimation, and some Robust and Resistant Alternatives

By P. J. GREEN

University of Durham, UK

[*Read before the* Royal Statistical Society *at a meeting organised by the* Research Section *on Wednesday, December 7th, 1983*, Professor J. B. Copas *in the Chair*]

SUMMARY

The scope of application of iteratively reweighted least squares to statistical estimation problems is considerably wider than is generally appreciated. It extends beyond the exponential-family-type generalized linear models to other distributions, to non-linear parameterizations, and to dependent observations. Various criteria for estimation other than maximum likelihood, including resistant alternatives, may be used. The algorithms are generally numerically stable, easily programmed without the aid of packages, and highly suited to interactive computation.

Keywords: NEWTON-RAPHSON; FISHER SCORING; GENERALIZED LINEAR MODELS; QUASI-LIKELIHOOD; ROBUST REGRESSION; RESISTANT REGRESSION; RESIDUALS

1. PRELIMINARIES

1.1. *An Introductory Example*

This paper is concerned with fitting regression relationships in probability models. We shall generally use likelihood-based methods, but will venture far from familiar Normal theory and linear models.

As a motivation for our discussion, let us consider the familiar example of logistic regression. We observe $y_1, y_2, \ldots, y_m$ which are assumed to be drawn independently from Binomial distributions with known indices $n_1, n_2, \ldots, n_m$. Covariates $\{x_{ij}, i = 1, 2, \ldots, m;\ j = 1, 2, \ldots, p\}$ are also available and it is postulated that $y_i \sim B(n_i, \{1 + \exp(-\Sigma x_{ij}\beta_j)\}^{-1})$, for parameters $\beta_1, \beta_2, \ldots, \beta_p$ whose values are to be estimated. The important ingredients of this example from the point of view of this paper are:

(A) a regression function $\boldsymbol{\eta} = \boldsymbol{\eta}(\boldsymbol{\beta})$, which here has the form $\eta_i = \{1 + \exp(-\Sigma x_{ij}\beta_j)\}^{-1}$; and
(B) a probability model, expressed as a log-likelihood function of $\boldsymbol{\eta}$, $L(\boldsymbol{\eta})$, which in this case is

$$L = \sum_{i=1}^{m} \{y_i \log \eta_i + (n_i - y_i) \log (1 - \eta_i)\}.$$

In common with most of the problems we shall consider in this paper notice that, in the usual application of this example,

(i) $\boldsymbol{\eta}$ has much larger dimension than $\boldsymbol{\beta}$,
(ii) the probability model (B) is largely unquestioned, except perhaps for some reference to goodness of fit, and
(iii) it is the form of the regression function (A) that is the focus of our attention.

Typically we would be interested in selecting covariates of importance, deciding the form of the regression function, and estimating the values of the β_js.

Present address: Department of Mathematical Sciences, The University, Durham DH1 3LE.

0035–9246–84–46149 $2.00

1.2. *General Formulation*

We consider a log-likelihood L, a function of an n-vector $\boldsymbol{\eta}$ of predictors. Typically n is equal to, or comparable with, the number of individual observations of which the likelihood forms the density or probability function, but we shall be concerned also with cases where individual observations are difficult to define, for example with one or more multinomial samples.

The predictor vector $\boldsymbol{\eta}$ is functionally dependent on the p-vector $\boldsymbol{\beta}$ of parameters of interest: p is typically much smaller than n. We base our inference on the function $\boldsymbol{\eta} = \boldsymbol{\eta}(\boldsymbol{\beta})$ by estimating the parameters $\boldsymbol{\beta}$, and deriving approximate confidence intervals and significance tests. Initially we shall consider only maximum likelihood estimates, and suppose that the model is sufficiently regular that we may restrict attention to the likelihood equations

$$\frac{\partial L}{\partial \boldsymbol{\beta}} = \mathbf{D}^{\mathrm{T}}\mathbf{u} = \mathbf{0} \tag{1}$$

where $\mathbf{u}$ is the n-vector $\{\partial L/\partial \boldsymbol{\eta}\}$, and $\mathbf{D}$ the $n \times p$ matrix $\{\partial \boldsymbol{\eta}/\partial \boldsymbol{\beta}\}$. The standard Newton-Raphson method for the iterative solution of (1) calls for evaluating $\mathbf{u}$, $\mathbf{D}$ and the second derivatives of L for an initial value of $\boldsymbol{\beta}$ and solving the linear equations

$$\left(\frac{-\partial^2 L}{\partial \boldsymbol{\beta}\boldsymbol{\beta}^{\mathrm{T}}}\right)(\boldsymbol{\beta}^* - \boldsymbol{\beta}) = \mathbf{D}^{\mathrm{T}}\mathbf{u} \tag{2}$$

for an updated estimate $\boldsymbol{\beta}^*$. This procedure is repeated until convergence. Equation (2) is derived from the first two terms of a Taylor series expansion for $\partial L/\partial \boldsymbol{\beta}$: for a log-likelihood quadratic in $\boldsymbol{\beta}$, the method converges in one step.

Commonly the second derivatives in (2) are replaced by an approximation. Note that

$$\frac{-\partial^2 L}{\partial \boldsymbol{\beta}\boldsymbol{\beta}^{\mathrm{T}}} = -\sum \frac{\partial L}{\partial \eta_i}\,\frac{\partial^2 \eta_i}{\partial \boldsymbol{\beta}\boldsymbol{\beta}^{\mathrm{T}}} - \left(\frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{\beta}}\right)^{\mathrm{T}} \frac{\partial^2 L}{\partial \boldsymbol{\eta}\boldsymbol{\eta}^{\mathrm{T}}} \left(\frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{\beta}}\right)$$

and we replace the terms on the right by their expectations (at the current parameter values $\boldsymbol{\eta}$). By the standard arguments:

$$E\left(\frac{\partial L}{\partial \eta_i}\right) = 0$$

$$E\left(\frac{-\partial^2 L}{\partial \boldsymbol{\eta}\boldsymbol{\eta}^{\mathrm{T}}}\right) = E\left(\frac{\partial L}{\partial \boldsymbol{\eta}}\left(\frac{\partial L}{\partial \boldsymbol{\eta}}\right)^{\mathrm{T}}\right) = \mathbf{A},$$

say, and with this approximation (essentially Fisher's scoring technique) (2) becomes

$$(\mathbf{D}^{\mathrm{T}}\mathbf{A}\mathbf{D})(\boldsymbol{\beta}^* - \boldsymbol{\beta}) = \mathbf{D}^{\mathrm{T}}\mathbf{u}. \tag{3}$$

We will assume that $\mathbf{D}$ is of full rank $\mathbf{p}$, and that $\mathbf{A}$ is positive definite throughout the parameter space: thus (3) is a non-singular $p \times p$ system of equations for $\boldsymbol{\beta}^*$.

Rather than handle their numerical solution directly, note that they have the form of normal equations for a weighted least squares regression: $\boldsymbol{\beta}^*$ solves

$$\text{minimize } (\mathbf{A}^{-1}\mathbf{u} + \mathbf{D}(\boldsymbol{\beta} - \boldsymbol{\beta}^*))^{\mathrm{T}}\,\mathbf{A}(\mathbf{A}^{-1}\mathbf{u} + \mathbf{D}(\boldsymbol{\beta} - \boldsymbol{\beta}^*)), \tag{4}$$

that is, it results from regressing $\mathbf{A}^{-1}\mathbf{u} + \mathbf{D}\boldsymbol{\beta}$ onto the columns of $\mathbf{D}$ using weight matrix $\mathbf{A}$. Thus we use an iteratively reweighted least squares (IRLS) algorithm (4) to implement the Newton-Raphson method with Fisher scoring (3), for an iterative solution to the likelihood equations (1). This treatment of the scoring method *via* least squares generalizes some very long-standing methods, and special cases are reviewed in the next Section.

Two common simplifications are that the model may be linearly parameterized, $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$ say, so that $\mathbf{D}$ is constant, or that L has the form $\Sigma L_i(\eta_i)$ (e.g. observations are independent) so that

A is diagonal.

IRLS algorithms also arise in inference based on the concept of quasi-likelihood, which was proposed by Wedderburn (1974) and extended to the multivariate case by McCullagh (1983). Suppose that the n-vector of observations $\mathbf{y}$ has $E(\mathbf{y}) = \boldsymbol{\eta}$ and $\operatorname{var}(\mathbf{y}) = \sigma^2 \mathbf{V}(\boldsymbol{\eta})$, where $\boldsymbol{\eta} = \boldsymbol{\eta}(\boldsymbol{\beta})$ as before, σ^2 is a scalar, and the matrix-valued function $\mathbf{V}(\cdot)$ is specified. The log-quasi-likelihood Q is defined as any function of $\boldsymbol{\eta}$ satisfying

$$\frac{\partial Q}{\partial \boldsymbol{\eta}} = \mathbf{V}^{-}(\boldsymbol{\eta})(\mathbf{y} - \boldsymbol{\eta}) \tag{5}$$

where $\mathbf{V}^{-}$ is a generalized inverse. We estimate $\boldsymbol{\beta}$ by solving the quasi-likelihood equations

$$\mathbf{0} = \frac{\partial Q}{\partial \boldsymbol{\beta}} = \frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{\beta}} \frac{\partial Q}{\partial \boldsymbol{\eta}} = \mathbf{D}^{\mathrm{T}}\mathbf{u},$$

say. Since $E(\partial Q/\partial \boldsymbol{\eta}) = 0$ and $E(-\partial^2 Q/\partial \boldsymbol{\eta}\boldsymbol{\eta}^{\mathrm{T}}) = \mathbf{V}^{-}$, the Newton-Raphson equations with expected second derivatives have the form (3) with $\mathbf{A} = \mathbf{V}^{-}$.

Questions of existence and uniqueness of the maximum likelihood estimates in various cases are discussed by Wedderburn (1976), Pratt (1981) and Burridge (1981). The large-sample theory that justifies likelihood-ratio tests and confidence intervals for parameters will be found in Cox and Hinkley (1974, p. 294–304) and McCullagh (1983). In particular, the asymptotic covariance matrix for the estimate of $\boldsymbol{\beta}$ is $\sigma^2 [E(-\partial^2 L/\partial \boldsymbol{\beta}\boldsymbol{\beta}^{\mathrm{T}})]^{-1} = \sigma^2 (\mathbf{D}^{\mathrm{T}}\mathbf{A}\mathbf{D})^{-1}$.

The important connection between such theoretical results and the numerical properties of IRLS is that both are justified by the approximate quadratic behaviour of the log-likelihood near its maximum. Thus it is reasonable that IRLS should work when maximum likelihood is relevant.

1.3. *History and Special Cases*

From the first, Fisher noted that maximum likelihood estimates would often require iterative calculation. In Fisher (1925), use of Newton's method was mentioned,

$$\theta^* = \theta + \frac{dL}{d\theta} \bigg/ \left(-\frac{d^2 L}{d\theta^2}\right)$$

either for one step only, or iterated to convergence. Implicitly he replaced the negative second derivative, the observed information, at each parameter value by its expectation assuming that value were the true one. This technique, and its multi-parameter generalization, became known as "Fisher's method of scoring for parameters", and was further discussed by Bailey (1961 Appendix 1), Kale (1961, 1962) and Edwards (1972).

Use of the scoring method in what we term regression problems seems to date from Fisher's contributed appendix to Bliss (1935). This paper was concerned with dosage-mortality curves, a quantal response problem as in Section 1.1, except for the use of the probit transformation in place of the logit. The relative merits of using observed or expected information were discussed by Garwood (1941), and the method has become more generally known from the various editions of Finney's book on Probit Analysis (1947, 1952, Appendix II).

Moore and Zeigler (1967) discussed these binomial problems with an arbitrary regression function, and demonstrated the quite general connection with non-linear least-squares regression. Nelder and Wedderburn (1972) introduced the class of generalized linear models to unify a number of linearly parameterized problems in exponential family distributions. These models are discussed in Section 3.2.

The important connection between the IRLS algorithm for maximum likelihood estimation and the Gauss-Newton method for least-squares fitting of non-linear regressions was further elucidated by Wedderburn (1974).

Jennrich and Moore (1975) considered maximum likelihood estimation in a more general exponential family than did Nelder and Wedderburn; their approach is similar to ours, except that the predictors $\boldsymbol{\eta}$ must be the expected values of the observations.

Important recent contributions have come from McCullagh (1983) and Jørgensen (1983), particularly regarding the treatment of dependent observations.

2. PRACTICALITIES

2.1. *Multinomial data*

Much of our discussion of the detailed properties of IRLS algorithms that follows can be motivated by examples with multinomial data.

First consider a single multinomial distribution with polynomially parameterized cell probabilities, such as often arises with data on gene frequencies. In the usual model for the human ABO blood system, A and B are alleles co-dominant to O, so that under random mating with gene frequencies of p, q and r for A, B and O, the probabilities for the phenotypes A, B, AB and O are $p^2 + 2pr$, $q^2 + 2qr$, $2pq$ and r^2. Both sets of frequencies sum to 1, and on removing this redundancy we obtain a regression problem with three predictors on 2 parameters, $\boldsymbol{\eta}$ being defined as the probabilities of the phenotypes, A, B and AB, and $\boldsymbol{\beta}$ taken as $(\log p, \log q)^{\mathrm{T}}$. This is obviously a *non-linear* regression, and the derivative matrix $\mathbf{D}$ is

$$\mathbf{D} = 2\begin{bmatrix} pr & -pq \\ -pq & qr \\ pq & pq \end{bmatrix}.$$

Given observed frequencies y_1, y_2, y_3, y_4 for A, B, AB and O, with $\Sigma y_i = n$, we have $u_i = y_i/\eta_i - y_4/\eta_4$ and $A_{ij} = n(\delta_{ij}/\eta_i - 1/\eta_4)$, where $\eta_4 = 1 - \eta_1 - \eta_2 - \eta_3$.

Incidentally, if η_4 retains its identity as a predictor, the regression becomes of order 4 x 2 but $\mathbf{A}$ is now diagonal: this is essentially the algorithm used by Jennrich and Moore (1975) for this problem, with the exception that they take $\boldsymbol{\beta} = (p, q)^{\mathrm{T}}$.

More generally, data in the form of R multinomial samples on the same set of S response categories often arise in the regression analysis of categorical data. We may arrange the data as a two-way table of counts $y_{rs}, r = 1, 2, \ldots, R$; $s = 1, 2, \ldots, S$, and the log-likelihood is essentially $L = \Sigma\,\Sigma y_{rs} \log p_{rs}$ where $\Sigma p_{rs} = 1$ for all r and the RS cell probabilities p_{rs} are suitably parameterized. There has been considerable interest in the case where the categories $1, 2, \ldots, S$ are ordered; see for example McCullagh (1980), Anderson and Phillips (1981) and Pratt (1981). Then we may be prepared to write

$$\sum_{i=1}^{s} p_{ri} = \Psi\left(\frac{\theta_s - \Sigma x_{rj}\beta_j}{\tau_r}\right)$$

for some given distribution function Ψ. Here the matrix (x_{rj}) represents covariate information, and the parameters are $-\infty = \theta_0 < \theta_1 < \ldots < \theta_S = +\infty, \beta_1, \beta_2, \ldots, \beta_p$ and $\tau_1, \tau_2, \ldots, \tau_R$, subject to some constraint such as $\Sigma \log \tau_r = 0$ to ensure identifiability. (Often, in fact, $\tau_r \equiv 1$.) Motivation for this model comes from considering the response categories as an arbitrary grouping of unobservable underlying data on a continuous scale. The θs are not usually of interest. This grouped continuous model also arises, of course, in the case of genuine grouped data in several samples from a location-and-scale family, in which case the θs are known.

In this situation we may define the predictors $\boldsymbol{\eta}$ so that either η_{rs} or $\Psi(\eta_{rs})$ is $\Sigma_{i=1}^{s} p_{ri}, r = 1, 2, \ldots, R; s = 1, 2, \ldots, S-1$. The former choice simplifies $\mathbf{u}$ and $\mathbf{A}$ and facilitates the comparison of several alternative distribution functions Ψ: the latter alternative gives $\mathbf{D}$ a simple form. With either parameterization, $\mathbf{A}$ is tridiagonal: this reduces the numerical algebra, and saves storage space. During the iteration, it is not usually necessary to insist on the ordering of the $\boldsymbol{\theta}$ parameters, as results of Pratt (1981) and Burridge (1981) show that provided all categories are observed, the likelihood will attain its maximum at a point where the inequalities are satisfied.

Use of models such as these for categorical data was motivated by the case where Ψ is a logistic function and $S = 2$. This is the usual linear logistic model for binary data (Cox, 1970). An alternative model for ordinal data with a similar motivation is Anderson's stereotype model, a special case of the models in Anderson (1984), which we may write as

$$p_{rs} = \exp\left(-\gamma_s - \phi_s \sum_j x_{rj}\beta_j\right) \Bigg/ \sum_t \exp\left(-\gamma_t - \phi_t \sum_j x_{rj}\beta_j\right)$$

where $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ and $\boldsymbol{\phi}$ are to be fitted, under the constraints $\gamma_S = 0$ and $1 = \phi_1 > \phi_2 > \ldots > \phi_S = 0$. Here, the simplest parameterization for IRLS is to use $\eta_{rs} = -\gamma_s - \phi_s \Sigma_j x_{rj}\beta_j$, $r = 1, 2, \ldots, R$; $s = 1, 2, \ldots, S - 1$. However, this model suffers from rank deficiency of $\mathbf{D}$ at $\boldsymbol{\beta} = \mathbf{0}$, and there will be consequent numerical difficulties, and problems in the large-sample theory. Further there is no guarantee that at the computed maximum, the ϕs will be correctly ordered.

2.2. *Example*

For a straightforward application of the grouped continuous multinomial model, consider the data in Table 1, relating school GCE A-level score to university degree performance for fifteen years' intake to a certain university first degree course. Clearly there may be heterogeneities here caused by changing standards with time, and it would have been preferable *not* to combine these data over the years. We examine the data for a possible linear relationship between the log-odds of attaining degree class s or better and A-level score. Here, $s = 1, 2, 3$ or 4 for degree I, II_1, II_2 and III. Parallel regression lines in this domain define the proportional odds model (see McCullagh, 1980) and are equivalent to a logistically distributed latent response variable, which is categorized into unknown intervals to provide the degree classification. Specifically, then, we suppose that the probability that a student with an A-level score of x_r attains degree class $i = 1, 2, 3, 4$ is p_{ri} such that $\Sigma_{i=1}^{s} p_{ri} = \Psi(\theta_s - \alpha x_r)$, where $\Psi(u) = (1 + e^{-u})^{-1}$. Defining $\eta_{rs} = \Sigma_{i=1}^{s} p_{ri}$: $s = 1, 2, 3, 4$, $r = 1, 2, \ldots, 10$, and $\boldsymbol{\beta} = (\theta_1, \theta_2, \theta_3, \theta_4, \alpha)^{\mathrm{T}}$, and using an unweighted least squares regression of the empirical log-odds ratio to provide initial estimates for $\boldsymbol{\beta}$, IRLS converged in 4 iterations to the maximum likelihood solution $\boldsymbol{\theta} = (-6.803, -5.177, -3.763, -2.096)^{\mathrm{T}}$, $\alpha = -0.3915$ (standard error 0.040). This fit gave a deviance (likelihood-ratio statistic against the saturated model) of 48.5 on 35 degrees of freedom: it is therefore probably adequate as a summary of the data. Treating A-level score as a factor rather than a quantitative covariate reduced the deviance to 36.5 (27 d.f.) so that based on a nominal significance test, no non-linearity is suggested. Use of the Normal distribution function for the latent response variable made practically no difference, while the Gumbel distribution (the "complementary-log-log" link, or proportional hazards model) fitted less well, whether degree classes were arranged in increasing or decreasing order.

TABLE 1
A-level score and degree classification

	Degree class					
Score	I	II(i)	II(ii)	III	*Pass*	*Total*
15	22	13	10	3	0	48
14	20	21	31	9	2	83
13	13	43	31	16	10	113
12	7	21	35	18	5	86
11	3	21	26	32	8	90
10	3	17	25	20	12	77
9	1	10	9	15	11	46
8	1	2	4	12	6	25
7	0	1	2	6	1	10
6	0	0	2	1	0	3

For this example, of course, the allegedly latent response variable has an immediate interpretation as an examination mark underlying the degree classification. We might then know the cutpoints $\{\theta_s\}$, or rather, allow an intercept and scale and write

$$\sum_{i=1}^{s} p_{ri} = \Psi\left(-\frac{t_s - \gamma - \delta x_r}{\sigma}\right),$$

where $\{t_s\}$ are known. For illustration, if the cutpoints $\mathbf{t}$ are taken as $(75, 60, 45, 30)^{\mathrm{T}}$ then the maximum likelihood estimates of γ, δ and σ are 8.773, 3.835 and 9.736, with a deviance of 51.2 (37 d.f.). Thus each A-level point is "worth" nearly 4 examination marks; note, however, from the magnitude of the estimate for σ that there is considerable variation about this regression line.

The approach is not, of course, limited to the case of a single covariate. For example, we have successfully used IRLS to fit both the grouped continuous and stereotype models to the back pain study data of Anderson and Philips (1981) in which there are 6 response categories and three categorical explanatory factors.

2.3. *Convergence*

General experience seems to be that choice of starting values for the parameter estimates is not particularly critical. Jennrich and Moore (1975) make this point quite strongly, though they are working only in an exponential family framework. In a model where there is a danger of multiple maxima, it is of course important to repeat the iterative process from several different points in the parameter space, in order to obtain more confidence that the true global maximum has been obtained.

In certain specialized problems, accurate starting values can be obtained by explicit formulae. For the ABO genetic example, Rao (1973, p. 371) gives formulae that render an iterative solution almost unnecessary. The effect of ignoring such formulae and using quite arbitrary starting points is illustrated in Fig. 1 for this example, using the O, A, B, AB frequencies: 202, 179, 35 and 6, as used by Thompson and Baker (1981). It will be seen that convergence is successfully obtained from nearly every admissible point, but that the iterations can be rather wild unless a little thought is used to provide a sensible initial estimate.

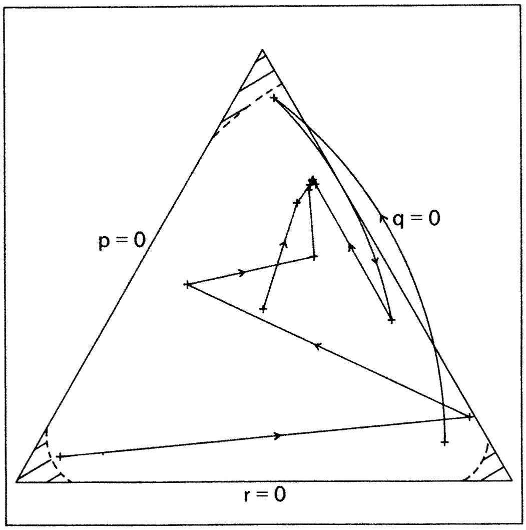

Fig. 1. The ABO blood group example: trajectories of successive iterations of IRLS from three different initial estimates, using the parameterization $\boldsymbol{\beta} = (\log p, \log q)^{\mathrm{T}}$, plotted in barycentric coordinates. The shaded region covers initial values for which IRLS is unsuccessful

For generalized linear models, it is easy to find usable starting values (See Section 3.2). In fact whenever the regression function is linear, $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$, and initial estimates of $\boldsymbol{\eta}$ can be obtained from a simple transformation of the data, an unweighted regression of these on the columns of **X** usually yields appropriate starting values of $\boldsymbol{\beta}$. Jørgensen (1983) suggests a modification for the non-linear case.

We have mentioned use of both the expected and observed information matrices in the iterative step of the algorithm. Again this decision is not usually important. Expected information used to be preferred because in the problems being considered it was algebraically simpler, and because its value on convergence was needed to compute the asymptotic variance of the estimates. For discussion of these points see Garwood (1941), Edwards (1972) and Cox and Hinkley (1974, p. 308), for example, but see also Efron and Hinkley (1978). Further, the observed information matrix may not be positive definite. In exponential family models appropriately parameterized, observed and expected information may be the same (Nelder and Wedderburn, 1972, Jennrich and Moore, 1975, see Section 3.2).

On the other hand, in certain situations the expected information is unknown (for example with censored survival data, the potential censoring times are not recorded for uncensored observations), or algebraically complicated, or involves nuisance parameters (see Section 3.1).

Only in simple cases can the behaviour of the algorithm on iteration be properly quantified. At worst, all we can say is that it is a fixed point method: if it converges to a point $\boldsymbol{\beta}$, then $\boldsymbol{\beta}$ is a solution of the likelihood equations. Exact Newton-Raphson is a quadratic method, so that convergence will be rapid near the solution, but may not be obtained at all far from this point. See Chambers (1977, p. 136) and Jennrich and Ralston (1979). As pointed out by Jørgensen (1983), we can always modify the Fisher scoring method by reducing the step size to ensure that the likelihood increases on each iteration.

The iterations must be monitored in order to detect convergence (or its failure). Two obvious methods for doing so are to record relative changes in parameter estimates or absolute changes in the log-likelihood. GLIM used a modification of the latter, but the former is more readily adapted to alternatives to likelihood methods (Section 5) and involves simpler calculations, while it is less suited to automatic application. When handling an unfamiliar problem, it is important to follow the entire iterative history of the solution to be confident that the convergence criterion employed is appropriate.

2.4. *Reparameterization*

One-to-one transformation of either $\boldsymbol{\eta}$ or $\boldsymbol{\beta}$ in the model $L = L(\boldsymbol{\eta}(\boldsymbol{\beta}))$ will not essentially change the problem, but does change its specification, and can make the difference between success and failure in the application of IRLS.

Suppose that $\boldsymbol{\eta}$ and $\boldsymbol{\beta}$ are in appropriately differentiable one-to-one correspondence with $\boldsymbol{\xi}$ and $\boldsymbol{\gamma}$ respectively, and denote the associated Jacobians by **S** and **R**. Thus **S** is $n \times n$ with $S_{ij} = \partial\eta_i/\partial\xi_j$, and **R** is $p \times p$ with $R_{ij} = \partial\beta_i/\partial\gamma_j$. If we re-parameterize the model as $L = L(\boldsymbol{\xi}(\boldsymbol{\gamma}))$ then by elementary calculus, **u**, **A** and **D** are replaced by $\mathbf{S}^{\mathrm{T}}\mathbf{u}$, $\mathbf{S}^{\mathrm{T}}\mathbf{AS}$ and $\mathbf{S}^{-1}\mathbf{DR}$. The likelihood equations $\mathbf{D}^{\mathrm{T}}\mathbf{u} = \mathbf{0}$ become $\mathbf{R}^{\mathrm{T}}\mathbf{D}^{\mathrm{T}}\mathbf{u} = \mathbf{0}$ and the IRLS step is

$$(\mathbf{S}^{-1}\mathbf{DR})^{\mathrm{T}}\,(\mathbf{S}^{\mathrm{T}}\mathbf{AS})\,(\mathbf{S}^{-1}\mathbf{DR})\,(\boldsymbol{\gamma}^* - \boldsymbol{\gamma}) = \mathbf{R}^{\mathrm{T}}\mathbf{D}^{\mathrm{T}}\mathbf{u},$$

which reduces to

$$(\mathbf{R}^{\mathrm{T}}\mathbf{D}^{\mathrm{T}}\mathbf{ADR})\,(\boldsymbol{\gamma}^* - \boldsymbol{\gamma}) = \mathbf{R}^{\mathrm{T}}\mathbf{D}^{\mathrm{T}}\mathbf{u}.$$

Thus, as expected, reparameterization in the $\boldsymbol{\eta}$-space has made no difference to the iterative solution, but in the $\boldsymbol{\beta}$-space it does make a difference, unless the transformation from $\boldsymbol{\beta}$ to $\boldsymbol{\gamma}$ is affine.

From a numerical point of view, there may be good reason to reparameterize in order approximately to linearize the problems. Successful use of IRLS depends essentially on the adequacy of approximate linearity of $\partial L/\partial \boldsymbol{\beta}$ in $\boldsymbol{\beta}$. If for some $\boldsymbol{\gamma}$, $\mathbf{R}^T\mathbf{D}^T\mathbf{u}$ is more nearly linear in $\boldsymbol{\gamma}$ than is $\mathbf{D}^T\mathbf{u}$ in $\boldsymbol{\beta}$, then transformation may be worthwhile.

For a simple example of this, consider again the ABO genetic system. If the problem were alternatively parameterized in terms of $\boldsymbol{\gamma} = (p, q)^T$, then the frequencies of phenotypes (A, B, AB) are $\boldsymbol{\eta}$ where $\boldsymbol{\eta}^T = (\gamma_1(2-\gamma_1-2\gamma_2), \gamma_2(2-2\gamma_1-\gamma_2), 2\gamma_1\gamma_2)$ and so

$$\mathbf{DR} = \frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{\gamma}} = 2\begin{bmatrix} 1-\gamma_1-\gamma_2 & -\gamma_1 \\ -\gamma_2 & 1-\gamma_1-\gamma_2 \\ \gamma_2 & \gamma_1 \end{bmatrix}.$$

Although this has a simpler form than does $\mathbf{D}$ of Section 2.1, it is found that with this set-up the problem is more sensitive to initial values for $\boldsymbol{\gamma}$. Figure 2 illustrates that starting values must be restricted to a much smaller part of the parameter space than with the earlier parameterization.

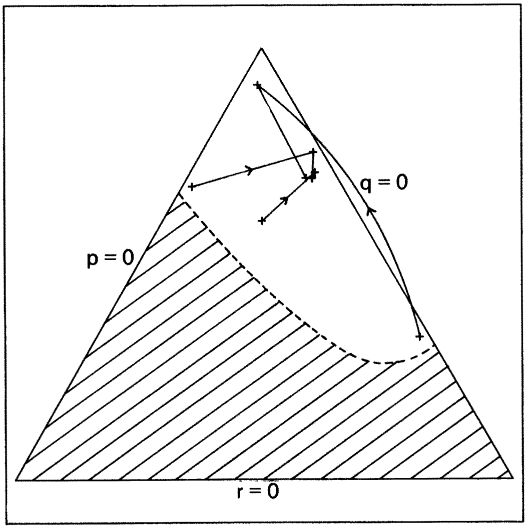

Fig. 2. As for Fig. 1, but with the alternative parameterization $\boldsymbol{\beta} = (p, q)^T$.

Although as we have seen, reparameterizing $\boldsymbol{\eta}$ will make no numerical difference (if we neglect rounding error) it may cause considerable changes in setting up the problem as the information matrix $\mathbf{S}^T\mathbf{A}\mathbf{S}$ may have simpler structure than $\mathbf{A}$.

2.5. *Computing the Weighted Least Squares Solutions*

The numerical analysis of the linear least squares problem is well developed. Chambers (1977, chap. 5) provides a useful summary from a statistical viewpoint. For ordinary least squares, finding $\boldsymbol{\beta}^*$ to minimize $\| \mathbf{y} - \mathbf{D}\boldsymbol{\beta}^* \| = (\mathbf{y}-\mathbf{D}\boldsymbol{\beta}^*)^T(\mathbf{y}-\mathbf{D}\boldsymbol{\beta}^*)$, there are methods that are more stable numerically than the obvious $\boldsymbol{\beta}^* = (\mathbf{D}^T\mathbf{D})^{-1}\mathbf{D}^T\mathbf{y}$.

The orthogonal decomposition methods involve the explicit or implicit construction of an $n \times n$ orthogonal matrix $\mathbf{Q}(\mathbf{Q}^T\mathbf{Q} = \mathbf{I})$ such that

$$\mathbf{QD} = \begin{bmatrix} \mathbf{R} \\ \mathbf{0} \end{bmatrix}$$

where $\mathbf{R}$ is a $p \times p$ upper triangular matrix. Since orthogonal transformations preserve euclidean length, our required solution $\boldsymbol{\beta}^*$ satisfies $\mathbf{R}\boldsymbol{\beta}^* = \tilde{\mathbf{Q}}\mathbf{y}$ (where $\tilde{\mathbf{Q}}$ denoted the first p rows of $\mathbf{Q}$)

and may be obtained by back-substitution. Note also that $\mathbf{D}^T\mathbf{D} = \mathbf{R}^T\mathbf{R}$, facilitating the calculation of $(\mathbf{D}^T\mathbf{D})^{-1}$ etc.

Orthogonal decompositions may be found by a modified Gram-Schmidt method, Givens' method, or by means of Householder transformations. This last approach is recommended for general use (see the procedures *decompose* and *solve* of Businger and Golub (1965), implemented as NAG routines F01AXF and F04ANF(NAG, 1981)).

From a strictly numerical-analytic point of view, the back-substitution for $\boldsymbol{\beta}^*$ should be followed by an iterative improvement, in which the residuals from the least squares solution become the right-hand sides for a new least squares problem. Providing the transformation has been saved this does not entail further decomposition; nevertheless, it is not likely to lead to any improvement relevant to data-analysis, particularly when the least squares solution is, as here, only a part of an iterative step.

Ordinary least squares solves our IRLS step (4) when $\mathbf{A}$ is a scalar matrix: the "modified dependent variate" $\mathbf{y}$ is $\mathbf{A}^{-1}\mathbf{u} + \mathbf{D}\boldsymbol{\beta}$. When $\mathbf{A}$ has a more complicated form, we need to use weighted least squares. Two possibilities are open to us:
(i) to transform the problem to ordinary least squares, or
(ii) to generalize the orthogonal decomposition method.
When $\mathbf{A}$ is diagonal, diag(v_i^2) say, the ttansformation (i) is trivial. With $\mathbf{y}$ defined as above, we minimize $(\mathbf{y} - \mathbf{D}\boldsymbol{\beta}^*)^T\mathbf{A}(\mathbf{y} - \mathbf{D}\boldsymbol{\beta}^*) = \Sigma v_i^2(y_i - \Sigma d_{ij}\beta_j^*)^2$ by component-wise multiplication of the entries in $\mathbf{y}$ and the rows of $\mathbf{D}$ by the square roots of the diagonal elements of $\mathbf{A}$, and using ordinary least squares. This gives the simple prescription:

$$\text{Regress } \frac{u_i}{v_i} + v_i\,\Sigma\, d_{ik}\beta_k \text{ on } \{v_i d_{ij}\} \tag{6}$$

Chambers (1977 p. 120) suggests this procedure, and there seems no point in looking for a method of type (ii).

For general, non-diagonal, $\mathbf{A}$, conversion to ordinary least squares entails construction of the Cholesky square root matrix $\mathbf{B}$ and using:

$$\text{Regress } (\mathbf{B}^{-1}\mathbf{u} + \mathbf{B}^T\mathbf{D}\,\boldsymbol{\beta}\,) \text{ on columns of } \mathbf{B}^T\mathbf{D} \tag{7}$$

An alternative approach would be to decompose $\mathbf{D}$ in the geometry determined by $\mathbf{A}$, by constructing an $n \times n$ matrix $\mathbf{Q}$ for which

$$\mathbf{Q}^T\mathbf{Q} = \mathbf{A},\ \mathbf{Q}\mathbf{D} = \begin{bmatrix} \mathbf{R} \\ \mathbf{0} \end{bmatrix}$$

and solving $\mathbf{R}\boldsymbol{\beta}^* = \tilde{\mathbf{Q}}\mathbf{y} = \tilde{\mathbf{Q}}(\mathbf{A}^{-1}\mathbf{u} + \mathbf{D}\boldsymbol{\beta}\,) = \tilde{\mathbf{Q}}\mathbf{A}^{-1}\mathbf{u} + \mathbf{R}\boldsymbol{\beta}$ as before, by back substitution.

This hybrid Householder/Cholesky decomposition may offer advantages in efficiency or accuracy, but we have experienced no difficulties with the routine use of a separate Cholesky decomposition followed by ordinary least squares. Note that $\mathbf{A}$ may be of special form (e.g. tridiagonal) which may be exploited in the decomposition.

With this approach, no special purpose software is needed for any of these computations. For example, the interactive general purpose language APL has all the array-handling capabilities that are required.

Of course, weighted least squares procedures are available in many statistical packages and subroutine libraries. In the BMDP series (Dixon, 1981), for example, the P3R and PAR programs fit non-linear regressions and the manual describes their usage for maximum likelihood estimation. The facilities in GLIM (Baker and Nelder, 1978), and GENSTAT (Alvey, *et al.*, 1977) for fitting generalized linear models will be reviewed in Section 3.3. Generalized least squares in which the weight matrix is not diagonal does not seem to be available in packages.

2.6. *Alternative Numerical Methods*

The Newton–Raphson method, whether or not the second derivatives are replaced by their expectations, is conceptually the simplest numerical procedure for maximizing the likelihood function. It would be naive to suggest that it can be used universally for these problems, but we have argued that one principal reason why it may break down (an inadequate approximation of L by a quadratic near its maximum) may also mean that maximum likelihood estimation is not appropriate.

However, two other reasons in particular may suggest that other numerical methods are more suitable—slow convergence far from the optimum, and the necessity of supplying analytic second derivatives. These difficulties can be overcome by use of Quasi-Newton methods (see for example, Gill and Murray (1972)). This has been recommended by Anderson and Philips (1981) and Anderson (1981) for the multinomial problems discussed in Section 2.1. Quasi-Newton methods seem highly suited to maximum likelihood estimation, and should perhaps be generally recommended if IRLS fails.

Other optimization methods that may be needed in particular applications include search methods (e.g. Powell (1964), Nelder and Mead (1965)) and the conjugate gradient algorithms (Fletcher and Reeves, 1964). These latter are economical in storage when the number of parameters is large, and are advocated by McIntosh (1982) for the fitting of linearly parameterized models on small computers. Search methods are probably most suited to problems where the objective function is rather less well behaved than most log-likelihoods.

Useful reviews of these methods are given by Chambers (1977, chapter 6), and Jennrich and Ralston (1979).

For certain problems, special purpose algorithms are available. For example, in log-linear models for multi-way contingency tables, one rival to the IRLS method is the iterative proportional fitting algorithm (see, for example, Bishop, Fienberg and Holland, 1976). This will normally be preferable, but in a sparse contingency table the difference is less clear. Brown and Fuchs (1983) provide a valuable discussion of these points.

The MLP package (Ross, 1980) uses several different algorithms, including a modified Newton method, for various special maximum likelihood problems, including probit analysis and genetic linkage. User-defined models can be analysed, and the program will handle non-linear parameterizations.

We should also touch here on the relationship to the *EM* algorithm (Dempster, Laird and Rubin, 1977). This is not a numerical method for maximum likelihood estimation in the same sense as IRLS. It is rather a general principle for handling problems in which the likelihood takes a particularly complicated form because of unobserved latent or missing variables. In particular cases there can be a close connection with IRLS. For example, Hinde (1982) shows that, for a Normal-Poisson compound distribution, when a numerical integral in the *EM*-algorithm is evaluated by Gaussian quadrature, the result is an IRLS algorithm. Brillinger and Preisler (1983) use a similar method for a variety of compound distributions.

3. NUISANCE PARAMETERS

3.1. *Introduction*

Not all of the parameters in a model need hold the same interest or have the same logical status. In certain cases, such "nuisance parameters" $\boldsymbol{\kappa}$ are additional to those naturally entering the regression function $\boldsymbol{\eta}$ $(\boldsymbol{\beta})$, and that is the situation considered here. (It is a different distinction, for example, than that made between treatments and blocks in a designed experiment.) It is generally profitable to recast the model as $L = L(\boldsymbol{\eta}, \boldsymbol{\kappa})$ where $\boldsymbol{\eta} = \boldsymbol{\eta}(\boldsymbol{\beta})$.

There are now two sets of likelihood equations

$$\frac{\partial L}{\partial \boldsymbol{\beta}} = \mathbf{D}^{\mathrm{T}}\mathbf{u} = \mathbf{0} \tag{8}$$

$$\frac{\partial L}{\partial \boldsymbol{\kappa}} = \mathbf{0}. \tag{9}$$

We continue to use the first set, but by analogy with the case of the variance in Normal linear regression we may replace (9) by some other criterion if convenient. We may in fact not wish to estimate $\boldsymbol{\kappa}$, but need to do so because of its implications for $\boldsymbol{\beta}$; solution of (8) may depend on $\boldsymbol{\kappa}$, and estimates of the variance of $\boldsymbol{\beta}$ may involve $\boldsymbol{\kappa}$.

In some situations, the dependence on $\boldsymbol{\kappa}$ may essentially factor out of the problem. Jϕrgensen (1983) discusses models which may be expressed in the form:

$$L = c(\mathbf{y}, \boldsymbol{\kappa}) + t(\mathbf{y}, \boldsymbol{\eta}) \;\; \phi(\boldsymbol{\kappa}) \tag{10}$$

where the vector $\mathbf{y}$ represents the observed data, and $\phi(\boldsymbol{\kappa})$ is a type of "precision" factor dependent on the nuisance parameters. Jϕrgensen calls this the extended class of generalized linear models, but in fact no linear structure $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$ is necessarily intended.

With the model (10), we have

$$\frac{\partial L}{\partial \boldsymbol{\beta}} = \phi(\boldsymbol{\kappa})\, \mathbf{D}^{\mathrm{T}} \frac{\partial t}{\partial \boldsymbol{\eta}}$$

and

$$-\frac{\partial^2 L}{\partial \boldsymbol{\beta}\boldsymbol{\beta}^{\mathrm{T}}} = -\phi(\boldsymbol{\kappa}) \left(\mathbf{D}^{\mathrm{T}} \frac{\partial^2 t}{\partial \boldsymbol{\eta}\boldsymbol{\eta}^{\mathrm{T}}} \mathbf{D} + \sum_i \frac{\partial t}{\partial \eta_i} \frac{\partial^2 \eta_i}{\partial \boldsymbol{\beta}\boldsymbol{\beta}^{\mathrm{T}}} \right) \tag{11}$$

The Newton-Raphson method for estimating $\boldsymbol{\beta}$ thus seems to involve $\boldsymbol{\kappa}$ only through the precision parameter ϕ, and indeed this appears to cancel from the two sides of (2). However, while the second term on the right in (11) disappears on taking expectations of the second derivatives, (or would vanish anyway if $\boldsymbol{\eta}$ were linear in $\boldsymbol{\beta}$), the expectation of $\partial^2 t/\partial \boldsymbol{\eta}\boldsymbol{\eta}^{\mathrm{T}}$ generally involves $\boldsymbol{\kappa}$. As Jϕrgensen observes, this will be the case unless (10) is an exponential family with $\boldsymbol{\eta}$ the minimal canonical parameter (which covers Nelder and Wedderburn's models). Jϕrgensen advocates the compromise of ignoring the second term of (11) but not taking expectations in the first term. This iterative method, which he terms the "linearization method", can be realised by taking $\mathbf{u} = \{\partial t/\partial \boldsymbol{\eta}\}$ and $\mathbf{A} = \{-\partial^2 t/\partial \boldsymbol{\eta}\boldsymbol{\eta}^{\mathrm{T}}\}$. This will allow estimation of $\boldsymbol{\beta}$ without interference from the nuisance parameter $\boldsymbol{\kappa}$. The new difficulty that may arise in thus using "observed" rather than "expected" information is that $\mathbf{A}$ may not be positive-definite (or even semi-definite), at some points in the $\boldsymbol{\eta}$-space. IRLS would then break down completely.

If so, then an alternative to treating $\boldsymbol{\beta}$ and $\boldsymbol{\kappa}$ in the same manner may be available if (9) is explicitly soluble for $\boldsymbol{\kappa}$ given fixed $\boldsymbol{\beta}$. Often $\boldsymbol{\kappa}$ is one-dimensional, and enters (9) in a simple way. If this is the case, we can attempt a solution for $(\boldsymbol{\beta}, \boldsymbol{\kappa})$ by means of a 2-part iteration:
(i) holding $\boldsymbol{\kappa}$ fixed, use IRLS to update $\boldsymbol{\beta}$;
(ii) holding $\boldsymbol{\beta}$ fixed, solve (9) for an updated $\boldsymbol{\kappa}$.
The convergence theory of this is even more inaccessible, but some results are available for the linear regression case (Section 4.1).

3.2. *Generalized Linear Models*

Nelder and Wedderburn (1972) proposed a class of likelihood functions for n independent observations $\{y_i\}$ which may be written

$$L = \sum_{i=1}^{n} \left(\phi\pi_i(y_i\theta_i - b(\theta_i)) + c(y_i, \phi) \right) \tag{12}$$

where $b(\cdot)$ and $c(\cdot)$ are prescribed functions, $\{\pi_i\}$ a set of known "prior weights", ϕ a nuisance precision parameter (known or unknown), and the canonical parameter θ_i is functionally related to the linear predictor $\eta_i = \Sigma\, x_{ij}\beta_j$.

This falls into our general framework, and in fact forms a subclass of Jφrgensen's extended generalized linear models. In the notation of Section 1, **D** is constant (the model or design matrix **X**), **A** is diagonal since the observations are independent, and **u** has a special form: because of the exponential family assumption (12) it involves only $y_i - b'(\theta_i) = y_i - E(y_i)$.

Standard examples include:

(i) $y_i \sim N(\eta_i, \sigma^2)$
(ii) $y_i \sim$ Poisson, with mean exp (η_i)
(iii) $y_i \sim$ Binomial, with probability $(1 + \exp(-\eta_i))^{-1}$
(iv) $y_i \sim$ Gamma, with mean η_i^{-1}
(v) $y_i \sim$ Binomial, with probability $\Phi(\eta_i)$
(vi) $y_i \sim$ Gamma, with mean η_i
(vii) $y_i \sim$ Negative binomial, with probability $(1 + \exp(-\eta_i))^{-1}$

where in each case $\eta_i = \Sigma\, x_{ij}\beta_j$. Thus many standard analyses, including linear regression, log-linear models and logit and probit analysis fit into this structure.

Using the IRLS approach described in Section 1 for generalized linear models has additional justification when the canonical parameter θ_i in (12) is identical to η_i. This includes the first four of the standard models listed above. If this holds then it is easy to see that the second derivatives of L with respect to $\boldsymbol{\eta}$ or $\boldsymbol{\beta}$ do not involve **y**: the observed and expected information matrices therefore agree, and we are using Newton–Raphson exactly. Statistically this property implies the existence of sufficient statistics: in the case where the prior weights are all unity and ϕ is known, we see that the likelihood L in (12) can be rearranged to exhibit $\mathbf{X}^T\mathbf{y}$ as sufficient for $\boldsymbol{\beta}$. For examples (v) to (vii), the observed and expected second derivatives differ: exact Newton–Raphson will work for (v) and (vii) but not (vi).

3.3. *GLIM and Composite Link Functions*

The unity of generalized linear models has been exploited in the development of the GLIM program (Baker and Nelder, 1978) which is essentially an IRLS algorithm coupled with data-handling facilities, automatic specification of several standard models, and a high-level syntax for manipulating design matrices. Similar features are now available in GENSTAT (Alvey, *et al.*, 1977). GLIM permits the fitting of non-standard ("user-defined") models by way of the OWN directive and its associated macros.

It may be shown that for a generalized linear model, after cancelling out the nuisance parameter ϕ as in Section 3.1, $u_i = \pi_i(y_i - \mu_i)/\tau_i^2\,\delta_i$ and $A_{ii} = \pi_i/\tau_i^2\,\delta_i^2$, where $\mu_i = E(y_i)$, $\tau_i^2 = \phi\pi_i\,\mathrm{var}\,(y_i)$ and $\delta_i = d\eta_i/d\mu_i$. Given a problem specified in terms of our **u**, **A** and **D**, it seems impossible for GLIM to handle non-diagonal **A**. If **A** is diagonal, the user's macros should define the GLIM vectors %YV, %FV, %VA, %DR and %PW ($\mathbf{y}$, $\boldsymbol{\mu}$, $\boldsymbol{\tau}^2$, $\boldsymbol{\delta}$ and $\boldsymbol{\pi}$) to satisfy the above relations for u_i and A_{ii}, and then use the columns of **D** in the FIT directive. Because of the redundancy in notation, there are several ways of setting up such a problem.

Considerable ingenuity has been expended in coding some rather unamenable problems into the GLIM command language following this prescription. In particular, Thompson and Baker (1981) demonstrated a useful extension of standard generalized linear models by way of composite link functions. A number of important models including the genetics example and the grouped continuous models that we have discussed fall into the distribution family (12) but with a non-linear regression function. Burn (1982) and Roger (1983) have extended this application of GLIM to a wide variety of multinomial problems arising in genetics.

The forthcoming replacement for GLIM will provide considerably better facilities for user-defined models, including composite link functions.

4. LINEAR REGRESSION

4.1. *Three IRLS Methods*

We now turn to the linear model

$$y_i = \Sigma\, x_{ij}\beta_j + \sigma\epsilon_i \tag{13}$$

in which the $\{\epsilon_i\}$ are independent and identically distributed with density f. We consider estimating $\boldsymbol{\beta}$ and σ according to one of three principles: (i) maximum likelihood, for known f; (ii) robust regression, providing estimates protected against departures of f from Normality; and (iii) resistant regression (which is deferred to Section 5.2).

Define ψ and w by $\psi(t) = tw(t) = -(d/dt)\log f(t)$, and write $\eta_i = \Sigma x_{ij}\beta_j$ and $r_i = (y_i - \eta_i)/\sigma$. The log-likelihood is $L = -n\log\sigma + \Sigma\log f(r_i)$, so the likelihood equations are

$$\frac{\partial L}{\partial\beta_j} = \sigma^{-1}\,\Sigma\,\psi(r_i)\,x_{ij} = 0 \tag{14}$$

and

$$\frac{\partial L}{\partial\sigma} = \sigma^{-1}(\Sigma\,\psi(r_i)\,r_i - n) = 0. \tag{15}$$

Letting asterisks denote updated items, if we substitute $w(r_i)\,(y_i - \eta_i^*)/\sigma^*$ for $\psi(r_i)$ then (15) and (14) yield

$$\sigma^{*2} = n^{-1}\;\Sigma\, w(r_i)\,(y_i - \eta_i^*)^2 \tag{16}$$

and

$$\Sigma w(r_i)\, y_i x_{ij} = \Sigma\Sigma w(r_i)\, x_{ij}x_{ik}\beta_k^*. \tag{17}$$

Of these, (16) updates σ^* from $\boldsymbol{\eta}^*$ explicitly, while (17) are normal equations for a weighted least squares regression, obtained without any appeal to the Newton-Raphson procedure.

Proceeding more formally, and differentiating (14), we derive the Newton-Raphson equations

$$\sum_i \frac{\psi(r_i)\,x_{ij}}{\sigma} = \sum_i\sum_k \frac{\psi'(r_i)\,x_{ij}x_{ik}}{\sigma^2}\,(\beta_k^* - \beta_k). \tag{18}$$

These least squares equations may be used directly, or $\psi'(r_i)$ can be replaced by one of two approximations. Firstly, noting that $\psi'(r) = w(r) + rw'(r)$, we can just ignore the second term (note that $w(\cdot) =$ constant for a Normal distribution). In this case (18) reduces to (17). Alternatively, Fisher scoring uses

$$\psi'(r) \simeq E(\psi'(r)) = \int \frac{\{f'(r)\}^2\,dr}{f(r)} = \alpha,$$

say, known as the intrinsic accuracy of the distribution (Fisher, 1925), so that (18) is replaced by

$$\sum_i \frac{\psi(r_i)\,x_{ij}}{\sigma} = \frac{\alpha}{\sigma^2}\sum_i\sum_k x_{ij}x_{ik}(\beta_k^* - \beta_k). \tag{19}$$

We have derived three alternative IRLS procedures for updating $\boldsymbol{\beta}$, based on (17), (18) and (19) respectively:

(I) Regress $\surd(w(r_i))y_i$ on $\surd(w(r_i))\,x_{ij}$, (17′)

(II) Regress $\dfrac{\sigma\psi(r_i) + \eta_i\,\psi'(r_i)}{\surd\psi'(r_i)}$ on $\surd(\psi'(r_i))\,x_{ij}$; (18′)

(III) Regress $\eta_i + \dfrac{w(r_i)}{\alpha}(y_i - \eta_i)$ on x_{ij}. (19′)

Of these procedures, (I) was used by Beaton and Tukey (1974), (II) is exact Newton-Raphson, and (III) is similar to that used by Huber (1973) and Bickel (1975), with the exception that they used an empirical estimate of α, in the absence of assumptions on f.

The three procedures coincide in the Normal case. In general, Newton's method (II) has the strongest justification, but it is usually more complicated to program, and is only defined when $\psi'(r) > 0$ for all r: that is, when ψ is strictly increasing (f, log-concave). Methods (I) and (III) are about equally simple to program, and (III) has the advantage of using only the unweighted design matrix $\mathbf{X}$.

Maximum likelihood for the error distribution $f(t) = kc^{1/k} \exp(-c \mid t \mid^k)/(2\Gamma(k^{-1}))$ coincides with L_k regression, which minimizes $\Sigma \mid y_i - \Sigma x_{ij}\beta_j \mid^k$. These models, and asymmetric versions in which the exponent multiplier depends on the sign of t are the only regressions in the class of models of Section 3.1: numerical solution is in principle easier since σ factors out of (14), and (15) has an explicit solution for σ. However, these models are notoriously difficult to fit when $k \leqslant 1$. IRLS cannot be recommended when $k < 2$: separate methods based on linear programming are available for L_1 regression (Barrodale and Roberts, 1973).

Dempster, Laird and Rubin (1980) discuss maximum likelihood regression, using method (I), for error distributions from the "Normal/independent" family, which includes the t distributions. They demonstrate that the IRLS algorithm so implemented is an example of an *EM* algorithm, so that theoretical results about convergence are available here.

Standard errors for the regression coefficients are readily obtained. We noted above that the expected negative second derivatives with respect to $\boldsymbol{\beta}$ are

$$E\left[-\frac{\partial^2 L}{\partial \boldsymbol{\beta}\boldsymbol{\beta}^{\mathrm{T}}}\right] = \frac{\alpha}{\sigma^2}\mathbf{X}^{\mathrm{T}}\mathbf{X}$$

so that the estimated asymptotic covariance matrix for the estimates of $\boldsymbol{\beta}$ is $\alpha^{-1}\hat{\sigma}^2(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$.

Turning now to robust regression, the principle of *M*-estimation (Huber, 1973) suggests estimating $\boldsymbol{\beta}$ by minimizing $\Sigma\,\rho((y_i - \Sigma x_{ij}\beta_j)/\sigma)$ for a suitably chosen loss function ρ. If ρ is $-\log f$ for some density function f then this is numerically equivalent to maximum likelihood again, assuming that f is the correct density.

We proceed as above using $\psi = \rho'$, except that in method (III) the intrinsic accuracy α must be replaced by an empirical estimate. Essentially the only other differences are in the basis for choice of the function ψ (from robustness considerations rather than a probability model), and in the treatment of σ. In robust regression it is usual to use criteria other than the maximum likelihood equation (15), and often, not to iterate on scale.

Holland and Welsch (1977) provide a useful discussion of these points, and of the choice of ψ function. No convergence theory is known for iterating on scale unless the Huber function $\psi(r) = \text{sign}(r)\min\{\mid r\mid, H\}$ (Huber, 1973) is used. They recommend using this first, and then, if a different ψ function is preferred, continuing without further change in σ. They compare eight different ψ functions in their paper, some based on likelihood models, and evaluate efficiencies and robustness properties by numerical integration and simulation.

4.2. *A Regression Example from Materials Science*

Delayed fracture in brittle materials may be demonstrated by observing the change in bend strength over a range of constant stress rates to failure. Braiden, Green and Wright (1982) use the classical Weibull model for distribution of brittle strength to derive a failure model in which y, the logarithm of the fracture stress, turns out to be related to x, the logarithm of the stress rate, by the linear regression $y = \beta_1 + \beta_2 x + \sigma\epsilon$. The error ϵ has a Gumbel distribution, and β_2 and σ are simply related to the stress corrosion and brittle fracture parameters that are of primary interest.

Raw data from one particular experiment is presented in Table 2. Note that the experiment was conducted at five different stress rates, with twelve independent observations at each, that there is considerable random variation in the data, and that a Normal linear regression, even after taking logarithms, would have suggested that there are several outliers. Each of the three IRLS methods of the previous section converges quickly to the maximum likelihood estimates of β_1, β_2 and σ: the resulting estimates appear in Table 3. For each of the methods, 7 iterations were needed to obtain the three estimates to relative accuracies of 10^{-5}, 10^{-4} and 10^{-3} respectively, and 11 iterations to obtain all three to 10^{-5}. There is nothing to choose between the methods on grounds of performance.

TABLE 2

Stress fatigue data. Double torsion test on specimens of tungsten carbide 6% cobalt alloy with ground surface finish. Fracture stress (MN m^{-2}) *at five different stress rates*

Stress rate (Mn m^{-2} s^{-1})				
0.1	1	10	100	1000
1676	1895	2271	1997	2540
2213	1908	2357	2068	2544
2283	2178	2458	2076	2606
2297	2299	2536	2325	2690
2320	2381	2705	2384	2863
2412	2422	2783	2752	3007
2491	2441	2790	2799	3024
2527	2458	2827	2845	3068
2599	2476	2837	2899	3126
2693	2528	2875	2922	3156
2804	2560	2887	3098	3176
2861	2970	2899	3162	3685

See Braiden, Green and Wright (1982).

TABLE 3

Linear regressions of the natural logarithms of the data in Table 2

	Estimates (s.e.'s in parentheses)		
	β_1	β_2	σ
Least squares	7.8089	0.021115	0.12887
Maximum likelihood:			
Ungrouped data	7.8667 (0.01675)	0.021867 (0.00420)	0.10594
Grouped data :			
Cut points 7.5(0.05)8.1	7.8663 (0.01654)	0.020454 (0.00408)	0.09947 (0.01052)
7.6(0.1)8.0	7.8657 (0.01678)	0.021717 (0.00459)	0.09662 (0.01198)

As an experiment, the same model was fitted after grouping the data, using the method for multinomial data outlined in Section 2.1. Remarkably, even with a very coarse grouping, the parameter estimates and their estimated standard errors are close to those for the ungrouped data. This is illustrated in Table 3 for two different groupings.

Braiden, Green and Wright (1982) gave further details of the analysis, which included a Monte-Carlo assessment of the adequacy of the Weibull model.

5. RESIDUALS AND RESISTANT METHODS

5.1. *Defining Residuals in Non-linear Models*

Residuals are traditionally thought of as (possibly standardized) discrepancies between observations and fitted values $\mathbf{y}-E(\mathbf{y})$. More recently, definitions have been sought which yield residuals uncorrelated under the assumed model and thus form a better basis for diagnostic determination of model inadequacy.

We have already met one such definition implicitly. Following the Householder decomposition for the simple linear model, we use the first p components of $\mathbf{Q}\mathbf{y}$ to determine $\boldsymbol{\beta}^*$ by back substitution. The remaining $(n-p)$ components are uncorrelated with variance σ^2, if the original data follows the given model and is uncorrelated with variance σ^2. Model inadequacy can thereby be diagnosed, but we cannot identify data inadequacy—the correspondence of "residuals" with "observations" has been lost.

Moving away from simple linear models, how are we to define useful "residuals" that can form a basis for assessment of model adequacy, detection of discrepant observations, and (to anticipate the next Section) accommodate possibly discrepant observations by their use in resistant analyses?

One basis for a general definition is to assign residuals not to the observations, but the predictors $\boldsymbol{\eta}$. The philosophy is that our model is prescribed by a likelihood function $L(\boldsymbol{\eta})$ where $\boldsymbol{\eta}$ varies *a priori* in an n-dimensional space. We then introduce restrictions on the freedom of $\boldsymbol{\eta}$ to vary by requiring $\boldsymbol{\eta} = \boldsymbol{\eta}(\boldsymbol{\beta})$, a given function of a p-vector $\boldsymbol{\beta}$ which is now the target of our inference. The $(n-p)$ degrees of freedom lost by thus parameterizing $\boldsymbol{\eta}$ force discrepancies between the data and the model $L(\boldsymbol{\eta}(\boldsymbol{\beta}))$. The residual assigned to each component of $\boldsymbol{\eta}$ should then measure the enforced change in $\boldsymbol{\eta}$.

A natural definition from this point of view would be to define a vector of residuals as $\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}(\hat{\boldsymbol{\beta}})$ where $\hat{\boldsymbol{\eta}}$ maximizes $L(\boldsymbol{\eta})$ and $\hat{\boldsymbol{\beta}}$ maximizes $L(\boldsymbol{\eta}(\boldsymbol{\beta}))$. This agrees with the usual definition for Normal linear regression $\mathbf{y} \sim N(\boldsymbol{\eta}, \sigma^2\mathbf{I})$; however, it is not invariant to trivial reparameterizations of $\boldsymbol{\eta}$. It is better to examine changes in $\boldsymbol{\eta}$ on the L-scale, so we define the n-vector of raw deviances $\boldsymbol{\Delta}$ by

$$\Delta_i = 2\left(\sup_t \{L(\boldsymbol{\eta}(\hat{\boldsymbol{\beta}}) + t\mathbf{e}_i)\} - L(\boldsymbol{\eta}(\hat{\boldsymbol{\beta}}))\right). \tag{20}$$

(where $\mathbf{e}_i$ is the unit vector in the ith direction), that is, Δ_i is twice the increase in log-likelihood attained by freeing η_i from its dependence on $\boldsymbol{\beta}$.

In the case of Normal linear regression, $\Delta_i = (y_i - \eta_i)^2/\sigma^2$, where η_i and σ^2 are the fitted mean and variance of y_i. More generally, if we can write

$$L(\boldsymbol{\eta}) = \sum_i L_i(\eta_i), \tag{21}$$

for example if we have n independent observations each parameterized by one η_i, then

$$\Delta_i = 2\left(\sup_{\eta_i'} (L_i(\eta_i')) - L_i(\eta_i)\right)$$

and we have the useful property that

$$\sum \Delta_i = \text{constant} - 2\,L(\boldsymbol{\eta}). \tag{22}$$

Thus in this case maximum likelihood is equivalent to minimum sum-of-deviances.

If (22) is deemed important, yet (21) does not hold, the only option seems to be to define deviances sequentially: for example, let

$$\Delta_i^* = \Delta_i^+ - \Delta_{i-1}^+, \text{ where } \Delta_i^+ = 2\left(\sup_{\substack{\boldsymbol{\eta}':\eta_j' = \eta_j \\ \text{for all } j > i}} \{L(\boldsymbol{\eta}')\} - L(\boldsymbol{\eta})\right) \tag{23}$$

Then the analogue of (22) holds for $\{\Delta_i^*\}$ and this definition is equivalent to (20) if (21) holds. A similar definition could be given for any ordering of the components of $\boldsymbol{\eta}$ in (23).

For elucidation of definition (23), suppose that L can be replaced by its approximating quadratic $L(\boldsymbol{\eta}) = c + \mathbf{b}^{\mathrm{T}}\boldsymbol{\eta} - \frac{1}{2}\boldsymbol{\eta}^{\mathrm{T}}\mathbf{A}\boldsymbol{\eta}$ where $c, \mathbf{b}$, and the non-negative definite matrix $\mathbf{A}$ are all constant. Then

$$\frac{\partial L}{\partial \boldsymbol{\eta}} = \mathbf{b} - \mathbf{A}\boldsymbol{\eta} = \mathbf{u} \text{ and } -\{\partial^2 L/\partial \boldsymbol{\eta}\boldsymbol{\eta}^{\mathrm{T}}\} = \mathbf{A}.$$

It may be shown that

$$\Delta_i^+ = \mathbf{u}_{(i)}^{\mathrm{T}}\,\mathbf{A}_{(ii)}^{-1}\,\mathbf{u}_{(i)} = \mathbf{z}_{(i)}^{\mathrm{T}}\,\mathbf{z}_{(i)}$$

where the subscripts in parentheses truncate after the ith row and column, $\mathbf{z} = \mathbf{B}^{-1}\mathbf{u}$, and $\mathbf{B}$ is the Cholesky square root of $\mathbf{A}$. Thus we have $\Delta_i^* = z_i^2$. This analysis is exact for the (correlated) Normal case, and otherwise only approximate.

For generalized linear models we have $\Delta_i = \Delta_i^* \simeq z_i^2$ where $z_i = \sqrt{\pi_i}(y_i - \mu_i)/\tau_i$ is just the ith observation standardized, after cancelling out the precision parameter ϕ; these are referred to as standardized residuals in the GLIM program and manual.

In general, note from (7) that the IRLS algorithm is seeking $\boldsymbol{\beta}$ such that $\mathbf{z} = \mathbf{B}^{-1}\mathbf{u}$ is uncorrelated with the columns of $\mathbf{B}^{\mathrm{T}}\mathbf{D}$. It is easy to calculate $\mathbf{z}$ (by forward substitution since $\mathbf{B}$ is lower-triangular), and the IRLS algorithm can be programmed to use these values directly for updating $\boldsymbol{\beta}$. Jϕrgensen (1983) has also suggested the use of $\mathbf{B}^{-1}\mathbf{u}$ as a vector of "score residuals". The above discussion suggests that they could be used interchangeably with the deviances.

However it is easy to find situations, even where the observations are independent, in which the use of score residuals does not make sense. In the case of linear regression (see Section 4.1) we have

$$u_i = \frac{\partial L}{\partial \eta_i} = \frac{\psi(r_i)}{\sigma}$$

$$A_{ii} = E\left(\frac{-\partial^2 L}{\partial \eta_i^2}\right) = \frac{\alpha}{\sigma^2},$$

so that, whether we use z_i as above, or without using expectations in the denominator, these residuals will not be defined and useful unless ψ is strictly increasing. This is precisely the condition for method (II) of Section 4.1 to work—that is, f must be log-concave.

This failure does not extend to the deviances—it is actually the quadratic approximation used above that breaks down. In fact, by (20) and (23), $\Delta_i = \Delta_i^* = 2\log(f_{\max}/f(r_i))$ where $f_{\max}$ is the supremum value attained by f. If f is bounded and continuous, this gives a sensible definition.

Residuals for diagnostic purposes in logistic regression have been discussed by Pregibon (1981). Of the various possible available scales, he finds Δ_i (or its signed square root) most useful, but also uses z_i. His paper, while nominally addressed to the binomial/logistic model, is relevant to all generalized linear models. Our discussion suggests that deviances should be of wider applicability, but that care is needed if independence (21) does not apply.

5.2. *Resistant Alternatives to Likelihood Methods*

The principle of resistance in statistical data analysis (as distinct from robustness) dictates that fitted models should be almost invariant to large changes in individual observations.

As applied to linear regression, this usually means that the least squares criterion:

$$\text{Choose } \boldsymbol{\beta} \text{ to minimize } \Sigma z_i^2 \tag{24}$$

where $z_i = z_i(\boldsymbol{\beta}) = y_i - \Sigma\, x_{ij}\beta_j$, is replaced by

$$\text{Choose } \boldsymbol{\beta} \text{ to minimize } \Sigma\, w(z_i)\, z_i^2 \tag{25}$$

where $w(z)$ is a weight function chosen to be 1 for small $|z|$ and declining as $|z|$ increases, in order to give less weight to "discrepant" observations (discrepant in the sense only of fitting the model less well). As implemented by Tukey (1977) and others (e.g. McNeil (1977)), resistant regression is achieved not through (25) but by

$$\text{Solving } \sum_i w(z_i)\, z_i x_{ij} = 0 \text{ for all } j. \tag{26}$$

Thus the normal equations, not the sum-of-squares criterion, are weighted. This distinction has led to some confusion. The obvious IRLS approach of updating $\boldsymbol{\beta}$ to $\boldsymbol{\beta}^*$ by regressing

$$\{w(z_i)\}^{\frac{1}{2}} y_i \text{ on } \{w(z_i)\}^{\frac{1}{2}} x_{ij} \text{ to get } \beta_j^* \tag{27}$$

converges, if at all, to a solution of (26), not (25). If it is really intended to solve (25), then differentiation leads to normal equations of form (26) but with $w(\cdot)$ replaced by a different weight $w^*(z) = w(z) + \frac{1}{2}\, zw'(z)$.

Unfortunately, for many sensible weight functions $w(\cdot)$, including Tukey's bi-square, $w(z) = (\max\{0, 1 - cz^2\})^2$, $w^*(\cdot)$ is not a valid weight function as it does not remain non-negative. Thus this IRLS approach will not apply for such $w(\cdot)$.

In this section we discuss the resistant methods obtained when (24) is regarded as maximum likelihood for Normal linear regression, and this model is replaced by an arbitrary one. In view of the points made above, it seems most natural to regard (26) rather than (25) as to be generalized.

At least in the case where $L(\boldsymbol{\eta}) = \Sigma_i L_i(\eta_i)$ we are led to consider weighted likelihood equations:

$$\sum_i w_i \frac{\partial L_i}{\partial \beta_j} = 0 \text{ for all } j \tag{28}$$

where the weights w_i depend on the discrepancy between the data and the fitted model. For measures of discrepancy, we usually use the deviances of the previous section. Pregibon (1982) discusses such resistant fits for the binomial/logistic model, again in terms of deviances, but specified by analogy with (25) as minimizing $\Sigma\lambda(\Delta_i)$ where $\lambda(\cdot)$ is a differentiable non-decreasing function. Again, this can be cast in the form (28).

However the weights are obtained, it seems natural to attempt to solve (28) by IRLS. In matrix notation we have $\mathbf{D}^{\mathrm{T}}\mathbf{W}\mathbf{u} = \mathbf{0}$ where $\mathbf{W} = \operatorname{diag}(w_i)$. In general, all of $\mathbf{u}$, $\mathbf{W}$ and $\mathbf{D}$ depend on $\boldsymbol{\beta}$, but in the spirit of our earlier discussion it is tempting to approximate the "second derivatives" by treating $\mathbf{W}$ and $\mathbf{D}$ as fixed.

Updated estimates are thus obtained from the equations $\mathbf{D}^{\mathrm{T}}\mathbf{WAD}(\boldsymbol{\beta}^* - \boldsymbol{\beta}) = \mathbf{D}^{\mathrm{T}}\mathbf{W}\mathbf{u}$, so that $\boldsymbol{\beta}^*$ is chosen to minimize

$$(\mathbf{Wu} + \mathbf{WAD}(\boldsymbol{\beta} - \boldsymbol{\beta}^*))^{\mathrm{T}}\, (\mathbf{WA})^{-1}\, (\mathbf{Wu} + \mathbf{WAD}(\boldsymbol{\beta} - \boldsymbol{\beta}^*))$$

$$= \sum_i w_i \left(\frac{u_i}{v_i} + v_i \sum_j d_{ij}\, (\beta_j - \beta_j^*) \right)^2 \tag{29}$$

Thus the only complication is an additional set of weights in the regression. Here of course w_i, as well as u_i, v_i and d_{ij} are calculated at the current value, β_j.

A program to fit the model conventionally may easily be modified to produce a resistant fit. It will be seen from (29) that it is necessary only to multiply u_i and v_i^2 by w_i immediately before the least squares step. In the case of generalized linear models this can be achieved using GLIM by either multiplying the prior weights $\pi_i (= \%\text{PW})$ by w_i, or dividing the variance function $\tau_i^2 (= \%\text{VA})$ by w_i. Using the second alternative an OWN fit may be specified. It will usually be necessary also to redefine the deviance terms %DI to give a sensible convergence criterion.

It is found that the presence of $\{w_i\}$ adversely affects the convergence properties of this IRLS method. If the process does converge, it will be to a solution of (28), but it may not converge at all. Some experimentation with different starting points, for example using L_1 rather than unweighted least squares initially, is often needed. It is occasionally necessary to start with constant weights ($w_i \equiv 1$), and gradually change them towards the desired values as the iterations progress. These suggestions may seem subjective and *ad hoc*, but it is not the aim of resistant data-analysis to provide a unique "objective" solution, but rather to examine whether any doubt should be cast on a model fitted conventionally.

When the likelihood is not of the form $\Sigma\, L_i(\eta_i)$, the only way of proceeding seems to be to replace

$$\text{Maximize } L = \text{Minimize } \Sigma\, \Delta_i^*$$

by

$$\text{Minimize } \Sigma\lambda(\Delta_i^*) = \text{Solve } \Sigma w_i \,\frac{\partial \Delta_i^*}{\partial \beta_j} = 0 \text{ for all } j.$$

Formally this can be treated just as above, although the problem is now not diagonal. We can proceed as if

$$u_r = \Sigma\, w_i \,\frac{\partial \Delta_i^*}{\partial \eta_r}\,, A_{rs} = \Sigma\, w_i \left(\frac{-\partial^2\, \Delta_i^*}{\partial \eta_r\, \partial \eta_s}\right)$$

but this method is untested. Note that the definitions of Δ_i^* assume an ordering on the components of $\boldsymbol{\eta}$, and so the behaviour of the algorithm, and its solution, may depend on this ordering.

5.3. *An Example from Probit Analysis*

For an illustration of a resistant analysis we consider the experiment described by Finney (1952, p. 69) which assessed the relative potency of three poisons. The data are given in Table 4.

TABLE 4
Relative potency of three poisons

Observation number	*Kill*	*Out of*	*Poison*	*Log dose*
1	44	50	R	1.01
2	42	49	R	0.89
3	24	46	R	0.71
4	16	48	R	0.58
5	6	50	R	0.41
6	48	48	D	1.70
7	47	50	D	1.61
8	47	49	D	1.48
9	34	48	D	1.31
10	18	48	D	1.00
11	16	49	D	0.71
12	48	50	M	1.40
13	43	46	M	1.31
14	38	48	M	1.18
15	27	46	M	1.00
16	22	46	M	0.71
17	7	47	M	0.40

Poisons: R – rotenone
D – deguelin
M – mixture

From Finney (1952), p. 69.

Following Finney, we perform a series of probit analyses in which it is assumed that the kill probability for a log-dose x of poison r ($r = 1, 2, 3$) is $\Phi(\alpha_r + \beta_r x)$ where Φ is the standard Normal distribution function. Of interest here is the possibility that the α's or the β's are equal. Deviances from maximum likelihood fits are given in Table 5: clearly there is ample evidence that the regression lines are neither the same nor parallel. Interpretation is easier if the lines are parallel, as this situation corresponds to constant relative potency at all levels of response. We therefore persevere with the four-parameter model $(\alpha_1, \alpha_2, \alpha_3, \beta)$ and attempt a resistant analysis. For comparison, three alternative weight functions were used, selected from those listed by Holland and Welsch (1977) for robust linear regression:

TABLE 5

Maximum likelihood analyses for Finney's data

*Number of parameters**	*Observations omitted*	*Deviance*	*Degrees of freedom*
2	–	70.8	15
4	–	30.3	13
6	–	20.1	11
4	2 11 15	14.4	10
4	11 14 15	13.7	10
4	11 16 17	7.7	10
2	11 16 17	67.7	12
6	11 16 17	7.1	8

* 2 : α, β; 4 : α_1, α_2, α_3, β; 6 : α_1, α_2, α_3, β_1, β_1, β_3.

Bi-square: $w(z) = (\max(0, 1 - (z/B)^2))^2$
Cauchy: $w(z) = (1 + (z/C)^2)^{-1}$
Huber: $w(z) = \min(1, H/|z|)$

In each case there is a tuning constant, taken as $+\infty$ for an ordinary likelihood analysis, and reduced for greater resistance. For residuals $\mathbf{z}$ we use the score residuals, which are simply the observations standardized by their fitted means and standard deviations for this exponential family model. Both least-squares and least-absolute-deviations fits were used to provide initial estimates. For each weight function the tuning constant was gradually reduced until stability was achieved; see Table 6. This procedure generated various possible groups of candidate observations to be labelled discrepant. Likelihood analyses were performed with these omitted

TABLE 6

Resistant analyses of four-parameter model for Finney's data

Weight function	*Tuning constant*	*Initial values* $(2 = L_2, 1 = L_1)$	*Weighted deviance*	*Observations severely down weighted*
B	9	1 or 2	28.2	–
B	6	1 or 2	25.5	11 2
B	4.5	1 or 2	17.5	11 16 2
B	3	2	5.3	11 17 16
B	3	1	5.9	11 14 15
C	6	2	28.1	–
C	4	2	25.8	11 2
C	2	2	17.7	11 2
C	1.4	1 or 2	9.3	11 17 16
H	2	2	29.3	–
H	1.5	2	26.0	11 2
H	1	1 or 2	19.2	11 2
H	0.5	2	9.8	11 2 15
H	0.4	1 or 2	7.8	11 2 15

(Table 5) revealing in particular that a dramatically better fit for the 4-parameter, parallel regression line, model is obtained if observations 11, 16 and 17 are neglected. There is no evidence to suggest that the 2 or 6 parameter models are preferable, having omitted these observations.

We therefore settle with the conclusion that all observations except numbers 11, 16 and 17 are consistent with the four-parameter model in which the kill probabilities are $\Phi(-2.673 + 3.906\,x)$, $\Phi(-4.366 + 3.906\,x)$ and $\Phi(-3.712 + 3.906\,x)$ – for Rotenone, Deguelin, and the mixture. Following inspection of the data, Finney omitted the same three observations from his analysis altogether, a course of action defended on the grounds that the chief interest here is in the behaviour of the poisons at high concentrations. He then fits parallel probit curves and his estimates are similar to ours.

Note that we would not advocate the use of resistant methods in order simply to reject inconvenient data: such discrepancies should normally be followed up with the experimenter. It is obviously unsatisfactory to have to discard three observations from 17. Note also that the result here is not just a better fit, but a better fit to a simpler and more interpretable model.

ACKNOWLEDGEMENTS

I am very grateful to Julian Besag and Allan Seheult for stimulation and encouragement during the course of this work, to Peter McCullagh, Bent Jørgensen and David Brillinger for helpful comments on an earlier version, to Professor H. Marsh for the degree classification data, and to the referees for valuable suggestions on presentation and a number of additional references.

REFERENCES

Alvey, N. G. *et al.* (1977) *The Genstat Manual*. Rothamsted Experimental Station.

Anderson, J. A., and Philips, P. R. (1981) Regression, discrimination, and measurement models for ordered categorical variables. *Appl. Statist.*, **30**, 22–31.

Anderson, J. A. (1984) Regression and ordered categorical variables (with Discussion) *J. R. Statist. Soc.* B, **46**, (in press).

Bailey, N. T. J. (1961) *Introduction to the mathematical theory of genetic linkage*. Oxford: University Press.

Baker, R. J. and Nelder, J. A. (1978) *The GLIM System, Release 3*. Oxford: Numerical Algorithms Group.

Barrodale, I., and Roberts, F. D. K. (1973) An improved algorithm for discrete L_1 approximation. *SIAM J. Numer. Anal.* **10**, 839–848.

Beaton, A. E., and Tukey, J. W. (1974) The fitting of power series. *Technometrics*, **16**, 147–185.

Bickel, P. J. (1975) One-step Huber estimates in the linear model. *J. Amer. Statist. Assoc.*, **70**, 428–434.

Bishop, Y. M. M., Fienberg, S. E. and Holland, P. W. (1976) *Discrete Multivariate analysis: theory and practice*. Cambridge, MA: MIT press.

Bliss, C. (1935) The calculation of the dosage-mortality curve. *Ann. Appl. Biol.*, **22**, 134–167.

Braiden, P., Green, P. J. and Wright, B. D. (1982) Quantitative analysis of delayed fracture observed in stress tests on brittle materials. *J. Material Sci.*, **17**, 3227–3234.

Brillinger, D. R. and Preisler, H. K. (1983) Maximum likelihood estimation in a latent variable problem. In *Studies in Econometrics, Time Series, and Multivariate Statistics*, 31–65 (S. Karlin, T. Amemiya and L. A. Goodman, eds.) New York: Academic Press.

Brown, M. B. and Fuchs, C. (1983) On maximum likelihood estimation in sparse contingency tables. *Computational Statistics and Data Analysis*, **1**, 3–15.

Burn, R. (1982) Loglinear models with composite link functions in genetics. In *Proceedings of the International Conference on Generalised Linear Models* p. 144–154. (R. Gilchrist, ed.) Lecture Notes in Statistics, vol 14. New York: Springer.

Burridge, J. (1981) A note on maximum likelihood estimation for regression models using grouped data. *J. Roy. Statist. Soc.* B, **43**, 41–45.

Businger, P. and Golub, G. H. (1965) Linear least squares solutions by Householder transformations. *Numer. Math.*, **7**, 269–276 (reprinted in Handbook for Automatic Computation, Vol. II, Linear Algebra, J. H. Wilkinson and C. Reinsch, eds. Berlin: Springer-Verlag).

Chambers, J. M. (1977) *Computational methods for data analysis*. New York: Wiley.

Cox, D. R. (1970) *Analysis of binary data*. London: Methuen.

Cox, D. R. and Hinkley, D. V. (1974) *Theoretical statistics*. London: Chapman and Hall.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977) Maximum likelihood from uncomplete data via the EM algorithm (with Discussion). *J. R. Statist. Soc.* B, **39**, 1–38.

——— (1980) Iteratively reweighted least squares for linear regression when errors are Normal/Independent distributed. In *Multivariate Analysis-V*. (P. R. Krishnaiah, ed.) Amsterdam: North Holland, p. 35–57.

Dixon, W. J. (ed.) (1981) *The BMDP Statistical Programs*. University of California Press.

Edwards, A. W. F. (1972) *Likelihood*. Cambridge: University Press.

Efron, B. and Hinkley, D. V. (1978) Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information. *Biometrika*, **65**, 457–487.

Finney, D. J. (1947, 1952) *Probit Analysis*. Cambridge: University Press.

Fisher, R. A. (1925) Theory of statistical estimation. *Proc. Camb. Phil. Soc.*, **22**, 700–725.

——— (1935) The case of zero survivors, In Bliss, (1935).

Fletcher, R. and Reeves, C. M. (1964) Function minimisation by conjugate gradients. *Computer J.*, **7**, 149–154.

Garwood, F. (1941) The application of maximum likelihood to dosage–mortality curves. *Biometrika*, **32**, 46–58.

Gentleman, W. M. (1973) Least squares computations by Givens transformations without square roots. *J. Inst. Maths. Applic.*, **12**, 329–336.

——— (1974) Basic procedures for large, sparse or weighted least-squares. *Appl. Statist.*, **23**, 448–454.

Gill, P. E. and Murray, W. (1972) Quasi-Newton methods for unconstrained optimisation. *J. Inst. Math. Applic.*, **9**, 91–108.

Golub, G. H. (1969) Matrix decompositions and statistical calculations. In *Statistical Computation*, 365–397. (R. C. Milton and J. A. Nelder, eds). New York: Academic Press.

Golub, G. H. and Styan, G. P. H. (1973) Numerical computations for univariate linear models. *J. Statist. Comput. Simul.*, **2**, 253–274.

Hinde, J. (1982) Compound Poisson regression models. In *Proceedings of the International Conference on Generalised Linear Models* 109–121. (R. Gilchrist, ed.) Lecture Notes in Statistics. Vol 14. New York: Springer.

Holland, P. W. and Welsch, R. E. (1977) Robust regression using iteratively reweighted least-squares. *Commun. Statist. Theor. Meth.*, **A6**, 813–827.

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

Jennrich, R. I. and Moore, R. H. (1975) Maximum likelihood estimation by means of non-linear least squares. *Proc. Amer. Statist. Assoc. (Statistical Computing Section)*, 57–65.

Jennrich, R. I. and Ralston, M. L. (1979) Fitting non-linear models to data. *Ann. Rev. Biophys. Bioeng.*, **8**, 195–238.

Jφrgensen, B. (1983) Maximum likelihood estimation and large-sample inference for generalised linear and non-linear regression models. *Biometrika*, **70**, 19–28.

Kale, B. K. (1961) On the solution of the likelihood equation by iteration processes. *Biometrika*, **48**, 452–456.

——— (1962) On the solution of likelihood equations by iteration processes. The multiparametric case. *Biometrika*, **49**, 479–486.

McCullagh, P. (1980) Regression models for ordinal data (with Discussion). *J. Roy. Statist. Soc.* B, **42**, 109–142.

——— (1983) Quasi-likelihood functions. *Ann. Statist.*, **11**, 59–67.

McIntosh, A. (1982) *Fitting Linear Models: An Application of Conjugate Gradient Algorithms*. Lecture notes in Statistics, vol. 10. New York: Springer.

McNeil, D. (1977) *Interactive data analysis*. New York: Wiley.

Moore, R. H. and Zeigler, R. K. (1967) The use of non-linear regression methods for analyzing sensitivity and quantal response data. *Biometrics*, **23**, 563–566.

NAG (1981) *Library Manual, Mark 8*. Oxford: Numerical Algorithms Group.

Nelder, J. A. and Mead, R. (1965) A simplex method for function minimisation. *Computer J.*, **7**, 308–313.

Nelder, J. A. and Wedderburn, R. W. M. (1982) Generalized linear models. *J. Roy. Statist. Soc.* A, **135**, 370–384.

Powell, M. J. D. (1964) An efficient method for finding the minimum of a function of several variables without calculating derivatives. *Computer J.*, **7**, 155–162.

Pratt, J. W. (1981) Concavity of the log likelihood. *J. Amer. Statist. Assoc.*, **76**, 103–106.

Pregibon, D. (1981) Logistic regression diagnostics. *Ann. Statist.*, **9**, 705–724.

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

Rao, C. R. (1973) *Linear Statistical Inference and its Applications*, 2nd ed. New York: Wiley.

Roger, J. H. (1982) Composite link functions with linear log link and Poisson error. *GLIM newsletter*, December 1982.

Ross, G. J. S. (1980) *Maximum Likelihood Program*. Rothamsted Experimental Station.

Thompson, R. and Baker, R. J. (1981) Composite link functions in generalised linear models. *Appl. Statist.*, **30**, 125–131.

Tukey, J. W. (1977) *Exploratory Data Analysis*. Addison-Wesley. Reading, Mass.

Wedderburn, R. W. M. (1974) Quasi-likelihood functions, generalised linear models, and the Gauss-Newton method. *Biometrika*, **61**, 439–447.

——— (1976) On the existence and uniqueness of the maximum likelihood estimates for certain generalised linear models. *Biometrika*, **63**, 27–32.

DISCUSSION OF DR GREEN'S PAPER

Dr Bent Jørgensen (Odense University): I am glad to see tonight's paper, and I find the paper difficult to criticize, because it promotes ideas that I have also been suggesting recently. My own starting point was an attempt to understand how GLIM fits generalized linear models. It is not so easy to understand Nelder and Wedderburn's (1972) paper, but it helps when one realizes that Fisher's scoring method may be interpreted as an iterative weighted least squares procedure for any type of distribution, not just for the exponential family. I wonder what led Dr Green on the track?

The iterative methods considered in the paper are all essentially of the form

$$\boldsymbol{\beta}^* = \boldsymbol{\beta} + \delta \mathbf{b},$$

$$\mathbf{b} = (\mathbf{D}^{\mathrm{T}}\mathbf{A}\mathbf{D})^{-1}\, \mathbf{D}^{\mathrm{T}} \partial L/\partial \boldsymbol{\eta}. \tag{a}$$

where $\mathbf{D} = \partial \boldsymbol{\eta}/\partial \boldsymbol{\beta}^{\mathrm{T}}$ and $\mathbf{A}$ is a positive-definite symmetric matrix. Since $\mathbf{D}$ has full rank, $\mathbf{D}^{\mathrm{T}}\mathbf{A}\mathbf{D}$ is positive-definite. Hence (a) is a gradient method, and the steplength $\delta \geqslant 0$ may be chosen to give an increase in the value of the likelihood compared with the previous iteration (Kennedy and Gentle, 1980, p. 430).

Although the computations in (a) may be performed via least-squares methods, I would like to suggest that the term iteratively reweighted least squares should only be used in connection with exponential families or quasi-likelihood estimation. Otherwise, I find the term too diffuse and ambiguous. I call (a) the delta-algorithm with weight matrix $\mathbf{A}$, a reminder of the similarity between the form $\mathbf{D}^{\mathrm{T}}\mathbf{A}\mathbf{D}$ and the way asymptotic variance matrices are transformed by the δ-method.

I would like to stress the importance of calculating the steplength properly, because this often makes the difference between success and failure of gradient methods, even for the Newton–Raphson algorithm used with a concave log-likelihood. In Section 2.6 it is asserted that the convergence of the algorithm is slow far from the optimum, but if the steplength is properly computed, the convergence is in fact quick far from the optimum, and the choice of starting value is not important. But the convergence may be slow near the optimum, because the matrix $-\mathbf{D}^{\mathrm{T}}\mathbf{A}\mathbf{D}$ may be a poor approximation to the second derivative of L, particularly if $\mathbf{D}$ is non-constant.

For the resistant estimating equation (28) there is no objective function, so the steplength can not be calculated in the usual way. It may be useful to define δ as the smallest positive solution to the equation

$$\mathbf{u}^{*\mathrm{T}}\mathbf{W}^*\mathbf{D}^*(\mathbf{D}^{\mathrm{T}}\mathbf{W}\mathbf{A}\mathbf{D})^{-1}\,\mathbf{D}^{\mathrm{T}}\mathbf{W}\mathbf{u},$$

where δ enters the equation through the updated quantities $\mathbf{u}^*$, $\mathbf{W}^*$ and $\mathbf{D}^*$. For maximum likelihood estimation, corresponding to $\mathbf{W} = \mathbf{I}$, this value of δ corresponds to the smallest local maximum of $L(\boldsymbol{\eta}(.))$ on the half line $\{\boldsymbol{\beta} + \delta\mathbf{b} : \delta \geqslant 0\}$.

As suggested in the paper we may take $\mathbf{A}$ as either the observed or expected information matrix for $\boldsymbol{\eta}$, but other choices for $\mathbf{A}$ are possible. If observed information is used when $L(\boldsymbol{\eta})$ is not concave, a simple modification of the Cholesky decomposition (Kennedy and Gentle, 1980, p. 445) may be used to obtain a positive-definite $\mathbf{A}$. If the log-likelihood is of the additive form (21), two possible choices for $\mathbf{A}$ are as a diagonal matrix with elements either

$$\mathbf{A}_{ii} = \frac{\partial L_i}{\partial \eta_i} \quad (\hat{\eta}_i - \eta_i) \qquad (i = 1, \ldots, n)$$

or

$$\mathbf{A}_{ii} = \left(\frac{\partial L_i}{\partial \eta_i}\right)^2 \Bigg/ \{2(L_i(\hat{\eta}_i) - L_i(\eta_i))\} \qquad (i = 1, \ldots, n),$$

where $\hat{\eta}_i$ maximizes $L_i(\eta_i)$. These choices are applicable if L_i is unimodal, and the first generalizes (17), whereas the second corresponds to an algorithm proposed by Ross (1982). I have used (17) to do L_1 regression in GLIM, and with a minor modification of $\mathbf{A}_{ii}$ for η_i near $\hat{\eta}_i$ this works well, and converges in less than twenty iterations. This and other exercises with the delta-algorithm suggest that by judicious choice of $\mathbf{A}$, the delta-algorithm may often cope with "notoriously difficult" problems. The choice of weight matrix for the delta-algorithm is discussed by Jørgensen (1983b).

The proper definition of residuals depends on the purpose for which they are intended, but as a rule there should be a one-to-one relation between residuals and observations, such that a large residual may be interpreted as an outlying observation. In general, the concept of an observation is somewhat elusive, but often we may parametrize the model in such a way that the components of $\boldsymbol{\eta}$ correspond to observations, in some sense. In the case $\mathbf{y} \sim N_n(\boldsymbol{\eta}, \boldsymbol{\Sigma})$, with $\boldsymbol{\Sigma}$ known, any definition should ideally produce $\mathbf{y} - \boldsymbol{\eta}$ as the vector of residuals, so let us try to apply the various definitions of residuals in Section 5.1 to this example. The score residuals give the vector of residuals $\mathbf{B}^{\mathrm{T}}(\mathbf{y} - \boldsymbol{\eta})$, where $\mathbf{B}$ is the square root of $\boldsymbol{\Sigma}^{-1}$, and hence fail the test, because the ith residual is a function of several $\mathbf{y}_i - \eta_i$. Similarly, (23) fails, whereas $\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}(\hat{\boldsymbol{\beta}}) = \mathbf{y} - \boldsymbol{\eta}(\hat{\boldsymbol{\beta}})$ are the correct residuals. For (20), it may be shown that a one-step approximation to determine the supremum, taking one step of Fisher's scoring method starting at $\boldsymbol{\eta}(\hat{\boldsymbol{\beta}})$, gives Δ_i as the square of the ith component of the vector

$$\mathbf{K}^{-1}(\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}\,\mathbf{D}(\mathbf{D}^{\mathrm{T}}\,\boldsymbol{\Sigma}^{-1}\mathbf{D})^{-1}\,\mathbf{D}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1})(\mathbf{y} - \boldsymbol{\eta}),$$

where $\mathbf{K}$ is diagonal. The main term of this expression is $\mathbf{K}^{-1}\,\boldsymbol{\Sigma}^{-1}(\mathbf{y} - \boldsymbol{\eta})$, so this definition also fails.

The definition I prefer is to take $\mathbf{d}^{1/2}\,\mathbf{A}^{-1}\,\mathbf{u}$ as the vector of residuals, where $\mathbf{d} = \mathrm{diag}\,\{A_{11}, \ldots, A_{nn}\}$, although this may be problematic if the log-likelihood is not concave. For the multivariate normal distribution this definition yields $(\mathbf{y}_i - \eta_i)/\Sigma_{ii}^{1/2}$ as the ith residual. Obviously, more work needs to be done on this subject.

My final comment is on the resistant fitting methods discussed in Section 5.2, and it probably reveals my ignorance on the subject. Apparently, resistant methods discard outlying observations from the fit, but according to the ultimate paragraph of the paper, this is not the purpose of the method. If one intends to find outlying or influential observations, one might instead use the regression diagnostic methods of Pregibon (1981). But then, what is the purpose of resistant methods?

Iterative techniques for maximum likelihood estimation resemble scientific investigation, in the sense that many small steps are taken in the direction of an unknown optimum. The difficulty with scientific investigation is that the objective function is not always clearly defined, and the weights attached to various issues vary from individual to individual. But I find that tonight's paper takes a good step in the right direction, and it is a pleasure for me to propose the vote of thanks.

Mr R. Thompson (Animal Breeding Research Organisation, Edinburgh): My first comments are related to the very generality of the approach, giving me difficulty in embedding specific problems into the framework. The colour inheritance of North Ronaldsay sheep (Ryder *et al.*, 1974) illustrates the point. Over 10 years ago I went through algebra similar to that in the ABO example, with a succession of models with over 100 terms to be differentiated and little structure in the equations. Later I realized that the models could be concisely represented by a loglinear model for the underlying genotypes and a composite matrix, **C**, linking the phenotypes to genotypes. The generalized linear model approach needed only slight modification using functions of the distribution as weights, and working dependent and regressor variables. In other examples the **C** matrix allows transformation from one scale allowing easy expression of the expectation of observations to another allowing independence of the observations, for example in Dr Green's Example 2.2. I hope that the next version of GLIM will allow the concise specification of **C** with model formulae, although the calculations are easily programmed in GENSTAT.

Again on the subject of generality I would like to have rules for parameterization for quick convergence. Why should $\log p$ be better than p in the ABO example? Although I seem to be the only person to have used them, the calculations for the speed of convergence of the *EM* algorithm (Dempster *et al.*, 1979) seem useful.

On the definition of residuals it seemed, at least in the written version, that the non-diagonal information matrix, **A**, was motivating the sequential definition of residuals. Estimation of parameters imposes correlation between residuals and this is taken account of in the cross-validatory definition (20). The definition of residuals should depend on the use intended for them and if **A** is non-diagonal then some statistic dependent on cross-products of suitable residuals might be informative. Can one expect the definition of residuals to be useful with binary data, when the residual deviance can be uninformative?

Given the confusion between (25) and (26) can the resistant methodology be interpreted as a formal model being a mixture of the data and the fitted model?

Over the years I can remember coming to hear important papers by Lindley and Smith, Curnow and Smith, Wilkinson *et al.* and Nelder and also managing to attend Smithfield Show. This year is no exception and I think Dr Green's paper will be as well-thumbed as these papers and so it gives me great pleasure to second the vote of thanks.

The vote of thanks was carried by acclamation.

Professor Murray Aitkin (University of Lancaster); This paper is a welcome addition to the small number of papers demonstrating the power and flexibility of IRLS approaches to maximum likelihood. As a confirmed GLIM user I will restrict my comments to three GLIM-related points:

(i) Censoring of observations presents an awkward problem for GLIM because of the different link function and (implicit) error model for the censored and uncensored observations. Composite link functions are complicated and not constructed' easily, if at all, in this case. Alternative maximum likelihood procedures using special features of the likelihood function work for right-censored observations from the Weibull and extreme value distributions (Aitkin and Clayton, 1980) and for the left- and right-censored observations from the logistic and log-logistic distributions (Bennett, 1983). For the normal and lognormal distributions the *EM* algorithm can be used (Aitkin, 1981), but the general IRLS approach described by the author is much simpler.

(ii) The discussion of robust and resistant methods is not reassuring. Without a probability model the "tuning constant" approach seems completely *ad hoc.* Model-based procedures are not difficult to construct, as the author notes in referring to the t-distribution method of Dempster *et al.* A likelihood-based "sensitivity analysis" equivalent to the tuning constant approach is possible by fitting the t-distribution for a fixed degrees of freedom k, and constructing a profile likelihood over k. This would clearly demonstrate the need, if it existed, for robust estimation.

(iii) The example in Section 5.3 seems to me an alarming example of the misuse of resistant methods. Dr Green has made it clear that this example is an illustration of *how* such methods fit into the IRLS approach, rather than an example of their value. On the probit scale, the regressions are not parallel for the complete set of data, so a simple comparison of relative potency of the three treatments is not possible. However the lines for Deguelin and Mixture are nearly parallel, and can be set equal with little change in deviance. Thus the relative potency of R to D or M depends on dose level, but that of D to M does not. The "discrepant observations" 11, 16 and 17, when removed, allow a common slope to be fitted. The simpler model is achieved at the expense of, not three observations, but 142. Surely this is bad modelling practice, even if one is interested only in the relative potencies at high dose levels.

Examination of the full probit model for the full data shows a bad fit of the model at points 10 and 11. Curvature of the Deguelin points on the probit scale is quite noticeable.

A complementary log-log link gives a considerable improvement, with a deviance of 16.09, instead of 20.13 for the 6-parameter model—and this is unchanged if the slopes for M and D are set equal. On the CLL scale there are no large residuals.

Surely it is better to present a model for the whole data which includes a necessary interaction, rather than removing the part of the data which establishes the need for the interaction and presenting a simpler model. To repeat, this is not a criticism of the paper, which I have found stimulating and valuable.

Mr R. Burn (Brighton Polytechnic): I would first like to comment on the use of IRLS for fitting multinomial models in genetics. Loglinear models with linear composite link functions (Thompson and Baker, 1981) which, as Dr Green points out, constitute a special case of his general class of models, are proving to be a convenient way of formulating multinomial models in which the cell probabilities are polynomial functions of the parameters to be estimated. With $\boldsymbol{\mu}$ denoting the vector of expected cell counts, the structure of these models is

$$\boldsymbol{\mu} = C\boldsymbol{\gamma}, \quad \boldsymbol{\gamma} = \exp(\boldsymbol{\eta}), \quad \boldsymbol{\eta} = X\boldsymbol{\beta},$$

where $\boldsymbol{\beta}$ is usually a vector of log gene frequencies or other genetic parameters. Although Dr Green's direct IRLS approach to such problems, exemplified by the ABO problem, is interesting, the composite link formulation appears to have at least two advantages. Firstly, there is no need to find analytic derivatives, which could require considerable effort for some problems. Secondly, the **C** matrix usually has some genetic meaning. In many cases **C** is an expression of the dominance relationships among the various alleles. For both of these reasons, together with the relative ease with which these models can be fitted in GLIM or GENSTAT, composite link models offer the prospect of a routine practical method for estimating genetic parameters. On the other hand, multinomial models do occur in which the cell probabilities are more complicated functions of the unknown parameters. There are problems, for instance, in which they are rational functions. Such models do not fit so naturally (if at all) into the composite link framework, and Dr Green's direct attack on the likelihood would be a useful alternative.

The other point that I would like to make concerns Dr Green's remark that packages such as GLIM or GENSTAT have no facilities for fitting models with a non-diagonal weight matrix. Although this appears to be true, it is still possible to implement the general algorithm, at least in GENSTAT, by first diagonalizing the matrix **A**. Since **A** is assumed positive definite and symmetric, there is an orthogonal matrix P such that $P^{\mathrm{T}}AP = \Lambda$, say, where Λ is the diagonal matrix of eigenvalues of **A**, which are positive. The IRLS step then becomes: regress $\Lambda^{-1}\,\tilde{\mathbf{u}} + \tilde{D}\boldsymbol{\beta}$ onto the columns of $\tilde{D}$ with weights Λ^{-1}, where $\tilde{\mathbf{u}} = P^{\mathrm{T}}\mathbf{u}$ and $\tilde{D} = P^{\mathrm{T}}D$.

I have tried this in GENSTAT on two of the examples in the paper, the ABO problem and the grouped continuous multinomial model for the A-level score data, with essentially the same results as the author. I would not suggest that this is necessarily an efficient way of computing IRLS solutions, but it may be convenient for those familiar with GENSTAT.

I would like to join the previous discussants in thanking Dr Green for a stimulating paper.

Dr Susan R. Wilson (Australian National University): Tonight's paper offers an overview of relevant literature on an important topic, namely the numerical evaluation of maximum likelihood estimates. The reasons for choosing one type of numerical method rather than another are complex, and some of these are outlined in Section 2.6. One further advantage of iterative least squares procedures is that they are more likely, than the alternative methods, to have convergence difficulties. Often closer inspection of the data will then reveal that the model is not at all appropriate for the data. One example where use of a quasi-Newton procedure led to an unacceptable model being adopted is given in Guerrero and Johnson (1982). The data were being used to predict the probability of female participation in the Mexican labour force as a function of dichotomous variables, locality, age, income and schooling. Guerrero and Johnson considered a model with the Box–Cox power transformation (with parameter λ) of the odds ratio as a linear function of the explanatory variables, thereby extending the logistic regression model ($\lambda = 0$). They concluded that $\lambda = -6{,}6$, and that the interaction between age and income was significant. The (correct) implementation of this extended model in GLIM (involving updating the model matrix on each iteration) fails to converge for these data. This enforces closer examination of the logistic regression model fit which reveals the reason. There is an aberrant observation, corresponding to "factor combination ac" (locality, income). Removal of this observation, and re-fitting the logistic regression model, a satisfactory fit is obtained, with now an interaction between locality and income in the linear part of the model. (Further details I have given elsewhere.) That their final model was inadequate should have been obvious from the parameter estimates and their corresponding standard errors (Guerrero and Johnson, 1982, Table 2). Also, other procedures, such as "resistant analysis" described in Section 5.2, and the use of graphical tools for determining the influence of individual observations on the power transformation, would alert one to the aberrant observation. In general, it should not be necessary to go so far in the overall analysis to find such a glaring discrepancy. To summarize, if IRLS fails, go back and examine the data carefully, *before* considering use of quasi-Newton methods.

Use of iterative least squares procedures for sparse contingency tables, where there are zeros in any of the marginal configurations defined by a log-linear model, is not well understood. In commenting on Brown and Fuchs (1983), Aston and Wilson (1984) describe how the implementation given by equation (4) cannot identify directly that some of the estimated cell frequencies for the model may be identically zero. Hence some of the parameter estimates and their standard errors will not be correct, although the fitted values will be correct. Stabilization of the parameter estimates and their standard errors can be readily achieved by identification of the occurrence of zeros in any of the marginal configurations, and then constraining the corresponding (zero) cells to have estimated cell frequencies of zero exactly.

Finally, I would like to add that the description of GLIM in Section 3.3 is a very limited one. In using GLIM's *own* directive and associated macros, I find the schematic diagram Fig. D1 (ignoring weights) helpful. Using a notation resembling that in tonight's paper, the diagram shows how the GLIM vectors (given in brackets) can be used, via the macros M1, M2, M3 and M4 to implement a generalized IRLS. Then it can be seen, that the GLIM notation is sufficiently flexible (not redundant), that by overwriting the GLIM vectors appropriately a wide variety [far wider than just the composite link function regression mentioned in Section 3.3 and the discussion] of problems can be solved in this framework. Although usually A, D, u, y and so on will correspond respectively to **A**, **D**, **u**, **y** they will not necessarily. For example, if **A** is not diagonal then a singular value decomposition can be used to produce a diagonal A with **D** being correspondingly altered to D and so forth. At which stage one abandons such manipulations and opts for writing a one-off program will depend on the individual. Most statisticians find it usually more efficient in terms of *their own* time to use a readily-available and familiar package. GLIM also has an OFFSET facility which enables iterative proportional fitting types of steps to be implemented, and so an even wider range of (conditional) maximum likelihood problems can be readily solved (further details are given in Adena and Wilson, 1982).

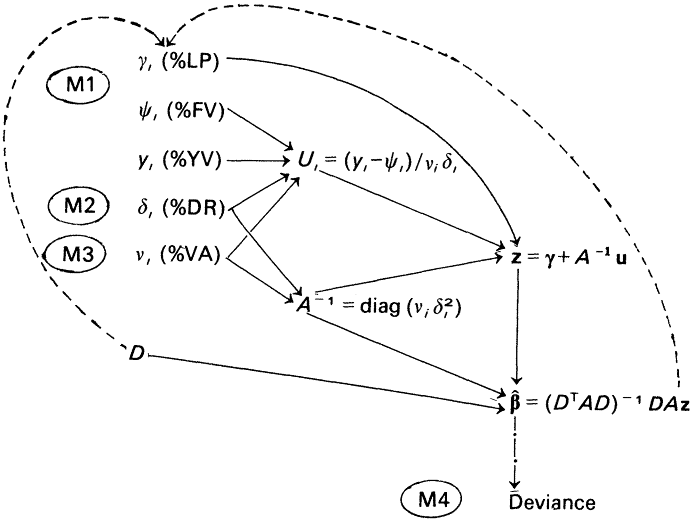
M1
γ_i (%LP)
ψ_i (%FV)
y_i (%YV)
M2
δ_i (%DR)
M3
v_i (%VA)
$U_i = (y_i - \psi_i)/v_i\delta_i$
$\mathbf{z} = \gamma + A^{-1}\mathbf{u}$
$A^{-1} = \text{diag}(v_i\delta_i^2)$
D
$\hat{\boldsymbol{\beta}} = (D^T AD)^{-1} DA\mathbf{z}$
M4
Deviance

Fig. D1

Dr J. A. Nelder (Rothamsted Experimental Station): First an etymological quibble: why "*re*weighted"? If you are iterating you can and do change the weight, but you may also be changing the dependent variate, the covariates, etc. I propose IWLS, not IRLS.

I agree with the proposer in finding sequentially defined residuals unsatisfactory; however, cross-validatory residuals can be defined for non-diagonal covariance structures, giving a (1, 1) correspondence between data values and residuals. They will of course be correlated, but so also are such residuals when the data are independent.

Professor Aitkin has commented on the fact that so-called resistant methods are not resistant to changes in the assumption about the link function. The same is true for variance functions; different variance functions will often lead to quite different subsets of points being given low weight. There are particular dangers in applying resistant methods to binary data, for if for a given unit the fitted value p is small then there are two possible residuals one small and negative if the datum value $y = 0$ and one large and positive if it is 1. In a substantial data set there will be a few ones, and these are highly informative about the size of p, yet a resistant analysis will "downweight" them drastically. I believe that "resistant" is a word that promises much more than it can deliver.

Finally, since the speaker has built his paper around an algorithm, could he not give us a classification of the models that can be fitted, characterized algorithmically, i.e. if some or all covariates change per iteration; similarly the dependent variate, weight, etc? This might be illuminating.

Mr G. J. S. Ross (Rothamsted Experimental Station): Dr Jϕrgensen has already referred to my use of what McCullagh and Nelder (1983) call "deviance residuals". These are quite simple to calculate, and enable the log likelihood to be expressed directly as a sum of squares when the error distribution is non-Normal (except in the unlikely case where the mode of the log likelihood is other than at $\boldsymbol{\eta} = y$). They may be differentiated, and used in any optimization algorithm for minimizing sums of squares, of which the Gauss–Newton method is the simplest. Details of the method are given in Ross (1982).

Direct comparison with IRLS may be made by solving the same problem by both methods from different initial parameter estimates. A graphical comparison in two dimensions is obtained by plotting contours of the log likelihood for the next iteration: large open contours indicate fast convergence. In all comparisons made so far the Gauss–Newton on deviance residuals is an improvement on IRLS. The reason is that the solution locus, when plotted in the space of deviance residuals, is more linear than in data space.

Further improvements may be achieved by stable parameter transformations, particularly with general non-linear models, but the improvements would also apply to IRLS. Stable transformations improve linearity and orthogonality within the solution locus.

As a contributor to statistical packages (MLP, GENSTAT) I would claim that it is preferable to use their optimization facilities rather than rely on the IRLS method, especially when the problem is unfamiliar. IRLS needs the correct algebraic specification, good initial estimates, but may still be slow or divergent. The optimization applications require only the model formula and the distribution of errors, there is protection from divergence, and parameter transformations are easily incorporated.

I cannot agree with Dr Wilson that the inefficiency of a method is a virtue if it shows data to be inadequate. There is no excuse for failing to interpret results fully.

I thank the author for a paper that has stimulated a useful discussion.

Professor D. M. Titterington (University of Glasgow): One approach to the approximate solution of equations (3) which has not been touched on in the paper is the recursive one. Let V_n^{-1} denote the matrix $D^{\mathrm{T}}AD$ based on n observations and suppose the observations are independent. Then $V_n^{-1} - V_{n-1}^{-1}$ is a matrix of rank 1 and V_n is easily computed from V_{n-1}. A sequence of recursive estimators for $\boldsymbol{\beta}$ can then be generated by

$$\tilde{\boldsymbol{\beta}}_n = \tilde{\boldsymbol{\beta}}_{n-1} + V_{n-1}\, \mathrm{S}(y_n, \tilde{\boldsymbol{\beta}}_{n-1}), \qquad n = 1, 2, \ldots,$$

where $\mathrm{S}(y, \boldsymbol{\beta})$ denotes the score vector for a single observation, y. Admittedly $\{\tilde{\boldsymbol{\beta}}_n\}$ usually depends on the order in which the observations are dealt with but the recursion is simple and sometimes reassuring asymptotic properties can be obtained via stochastic approximation theory. The case of linear logistic regression is dealt with by Walker and Duncan (1967); see also Goodwin and Payne (1977, Chapter 7). Titterington (1984) looks at recursive estimation based on incomplete data and also notes an interesting parallel between the general method of scoring and the *EM* algorithm. The former is defined by, say,

$$\hat{\boldsymbol{\beta}}_r = \hat{\boldsymbol{\beta}}_{r-1} + \{nI(\hat{\boldsymbol{\beta}}_{r-1})\}^{-1}\, \partial L(\hat{\boldsymbol{\beta}}_{r-1})/\partial \boldsymbol{\beta}, \qquad r = 1, 2, \ldots,$$

where $I(\boldsymbol{\beta})$ is the information matrix from one observation, typically hard to compute and invert in incomplete-data problems. An alternative iteration is obtained by using, instead of $I(\boldsymbol{\beta})$, the much simpler $I_c(\boldsymbol{\beta})$, corresponding to a complete observation. This iteration is sometimes exactly the *EM* algorithm and often very close to it.

Professor A. C. Atkinson (Imperial College, London): This interesting paper complements both the recent published and unpublished work of Jϕrgensen and also the book of McCullagh and Nelder (1983). My comments are however addressed mainly to Section 5.2 of the paper.

I remain unclear about the operational difference between robust and resistant procedures. In principle the distinction is clear, but, as Huber comments (1981, p. 7), the procedures are, for all practical purposes, the same. Even so, because of the effect of leverage, (25) or, not equivalently, (26), may not provide what is required. For observations at remote points in the space of the carriers, that is with $h_i = x_i^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}x_i$ approaching one, the least-squares residuals will be small, regardless of the observed value of the response. Such observations will not be down-weighted and the procedure will not be resistant. Procedures which do allow for the effect of leverage include the limited influence regression of Krasker and Welsch (1982), further discussed by Huber (1983). But there remain some problems of interpretation (Cook and Weisberg, 1983).

If there are outlying observations at points of high leverage, use of the author's robust/resistant procedure may lead to the down-weighting of observations with low leverage, some of which may have large residuals due to the distortion of the fitted model by the outliers. An example of this is provided by Ruppert and Carroll (1980) whose robust method was trimmed least squares. In their analysis of some data on salinity, the conclusion of one analysis was that observations 1, 13, 15 and 17 should be trimmed. I have elsewhere (Atkinson, 1982) contrasted this rather muddy conclusion with the results of a diagnostic analysis which clearly demonstrates that observation 16 has an extreme value of one of the carriers. It has however to be admitted that this value, although not its importance, can be detected by visual inspection of the marginal values of the explanatory variables.

Dr R. Gilchrist (Polytechnic of North London): It is a pleasure to add my congratulations to Peter Green on a thought-provoking paper. In the time allowed, I shall confine myself to discussing three aspects of the paper. Firstly, Dr Green is indeed correct that for "nuisance parameters", a two-stage algorithm can be more successful in achieving convergence than a single-stage updating algorithm. For example, in Scallan *et al.* (1984), it is shown how such an algorithm can be usefully employed in the exponential family framework to estimate what we call a *parametric link function*, i.e. where $E(Y_i) = \mu_i = h(\Sigma\, x_{ij}\beta_j, k)$, for some known function h. This works well, for example, for the Box–Cox link with unknown exponent or for estimating the shape parameter and asymptote of the generalized logistic curve defined by $\ln(\mu_i) = \ln(k_1) - k_2 \ln(1 + \exp(-\eta_i/k_2)$.

Moreover, this two-stage technique is easily incorporated into GLIM and is closely related to direct likelihood inference, as discussed by Aitkin (1982). For example, for one-dimensional k, our technique maximizes the likelihood by moving up the "profile likelihood". Dr Green is correct in asserting that GLIM 3 is not designed to handle non-diagonal **A**. However, as several other discussants have noted, it is *possible* to handle non-diagonal **A** in GLIM 3. For example, you can use a Cholesky decomposition as in Section 2.5 and incorporate this in the design matrix. Examples of this procedure for some standard time series models should shortly appear in the *GLIM Newsletter* (Scallan, 1984); however, this technique is clearly somewhat tricky for the non-specialist. It would not be difficult to change the GLIM code to allow non-diagonal **A**, although at the cost of some efficiency. However, the problem remains as to what form of **A** should be made available.

My final comments concern residuals. Firstly, orthogonal residuals have some appeal to me, despite their lack of independence for non-Normal models. For example, the Givens algorithm (available in GLIM 4) gives the so-called *recursive residuals* used by Brown *et al.* (1975). These can be useful for detecting changes in a model over time, even if they do not pinpoint discrepant observations. Secondly, the discussion in Section 5.1 may not make it crystal clear that the *standardised residuals* of GLIM 3 do not take into account the correlation between an observation y_i and its fitted value $\hat{\mu}_i$. Thus Gilchrist (1981) and Pregibon (1982) have suggested what Gilchrist refers to as *adjusted residuals*, which to some extent take into account this correlation. Moreover, Pregibon (1982) shows that it is these residuals which give the score test statistic corresponding to the hypothesis that a given observation is an outlier. (These adjusted residuals are trivially computed in GLIM 3 by dividing GLIM's standardized residuals by $\sqrt{(1-h_i)}$, where $h = \%WT^*\%VL/\%SC$.) Moreover, this technique can be applied in a similar way to the case where the observations are correlated. Finally, Dr Green's consideration of changes in η on the L-scale certainly seems valuable; however, I have the feeling that selected transformations of $\eta - \eta(\hat{\beta})$ may yield better tests against certain alternative hypotheses.

Mr A. C. Davison (Imperial College, London): I have a few remarks pertaining to the deviance residuals Δ_i defined at equation (20) of the paper.

Suppose that independent observations Y_i have common continuous distribution function $F(y; \eta, \alpha)$, predictors $\eta_i = \eta_i(\beta)$, and that α is a shape or scale parameter common to all the observations: σ in the Normal linear model, the shape parameter of the gamma or Weibull distributions, and so on.

In such a situation one would often work with the signed square roots

$$r_D(y_i; \hat{\eta}_i, \hat{\alpha}) = \operatorname{sgn}(t_{\max})\,\{\Delta_i\}^{1/2}$$

of the deviance residuals Δ_i, where $t_{\max}$ is the value of t for which the supremum in (20) is attained, and wish to treat them as approximately independent standard Normal variates. In particular, when the r_D are ordered and plotted against Normal order statistics, the resulting graph should be close to a straight line of unit slope through the origin if the r_D are to be useful in assessing closeness of fit of $F(.)$ to the data. How useful is this graph for a given distribution $F(.)$?

This problem may be studied by considering the function

$$G(w) = r_D\,(F^{-1}(\Phi(w); \eta, \alpha); \eta, \alpha),$$

which may depend on α, as a function of w. For

$$F(x; \eta, \alpha) = \Phi((x - \eta)/\alpha),$$

that is, the standard Normal-theory linear model,

$$r_D(y; \eta, \alpha) = (y - \eta)/\alpha,$$

$$F^{-1}(\Phi(w); \eta, \alpha) = \eta + \alpha w,$$

and so $G(w) = w$, as we would require of a sensible measure of closeness to Normality.

For the Weibull distribution, a Taylor expansion shows that the function $G(w)$ is to second order in w the expression $-0.345 + 1.023w + 0.0035w^2$ independently of the value of α: very close to the required straight line for values of w in the usual range but with a negative intercept.

For the generalized Pareto distribution

$$F(y; \eta, \alpha) = 1 - (1 - \alpha y/\eta)^{1/\alpha},$$

which is exponential when $\alpha = 0$, a distribution useful in some extreme-value contexts, the function $G(.)$ depends on the value of α but is quite close to the line $G(w) = w$ for the usual values of w and α.

Adjustments to bring the observed deviance residuals r_D close to the line $G(w) = w$ may sometimes be made; but it may be sufficient to know approximately the size and direction of the allowance the eye should make when inspecting the r_D.

The implication seems to be that appropriately defined deviance residuals for continuous distributions $F(.)$ often have properties very close to those of standardized residuals in the Normal-theory linear model, and therefore are likely to be easily interpretable and of potentially wide applicability, even for measuring how close to $F(.)$ the distributional form of the data is. The effect of the data actually having distribution $H(.)$ rather than $F(.)$ may be assessed in an obvious way. These remarks do not take into account the likely effect on the r_D of their mutual correlation: that may be quite a different story, and may perhaps be tackled by methods analogous to those of Cox and Snell (1968).

Professor T. Lewis (The Open University): Paraphrasing Vanzetti (Frankfurter and Jackson, 1929, p. xi), I wish to congratulate Dr Green on *most* of his paper. But Professor Aitkin's remarks on the probit example (Section 5.3) need following up.

Students wondering about what work to take up and what statisticians do are often advised to look at the journals and the RSS discussion papers. It would grieve me to think of them taking at face value Dr Green's words:

> ". . . clearly there is ample evidence that the regression lines are neither the same nor parallel. Interpretation is easier if the lines are parallel. . .. We therefore persevere with the four-parameter model . . ."

– in other words, bugger the data!

As Dr Green mentioned, the data in Table 4 come originally from Martin (1942). He feels a bit queasy about omitting the three points 11, 16, 17. Martin, the experimenter, who also omitted these points, talked about breaks in the regression lines, but in fact the deviations that bothered him are non-significant. Dr Green defends his omission of the three points by saying Finney did the same. So what? He cites Finney's argument (Finney, 1947, 1952) that only the behaviour at high concentration x is of interest. On this argument, the low-x points 3, 4, 5 should have been omitted along with 11, 16, 17. But they were conveniently retained.

In the third edition of his book (Finney, 1971, p. 233), Finney has a new text:

> "So far as fair interpretation of Martin's experiment is concerned, the rejection of anomalous observations may be questioned, but argument about this after so many years is unprofitable. Here the three low doses are omitted, solely for the purpose of illustrating numerical procedures."

Had Dr Green stated that his resistant analysis was solely for the purpose of illustrating numerical procedures, all would have been fine. But on the contrary, he claims

> ". . . that the result . . . is not just a better fit, but a better fit to a simpler and more interpretable model."

Professor Finney magisterially distances himself from the earlier analysis, as if it had nothing to do with him. Actually, if we go back to Martin's paper (Martin, 1942, pp. 69–81), we find it immediately followed by a companion paper by Finney (Finney, 1942, pp. 82–94). Mutual acknowledgements in the two papers suggest that Martin was the experimenter, Finney the statistical guru. Would Martin have thrown out his three points if his guru had advised against?

It is easy to criticize; the comeback is "Can you do better?" I do not know, but here is an attempt. First note points 8 and 7 as reminders of the magnitude of the standard errors attaching to the observations. Now look at the same-*x* points 1(*R*), 15(*M*), 10(*D*), at the closer group 3(*R*), 16(*M*), 11(*D*), and then at 17(*M*), 5(*R*), remembering that *M* is a mixture of *R* and *D* and not just any old poison. Are the lines *obliged* to be parallel? Might they not be *concurrent?* This would make sense, because at the concentration at which Rotenone and Deguelin have the same toxicity the mixture would plausibly also have this toxicity. Using all 17 points, I fitted the 5-parameter model

$$y_r = \Phi^{-1}(p_r) = \eta + \beta_r(x - \xi) \quad (r = 1, 2, 3).$$

The fitted point of concurrence $(\tilde{\xi}, \tilde{\eta})$ was (0.4,–1.2), the fitted slopes were $\tilde{\beta}_R = 4.17, \tilde{\beta}_M = 2.83, \tilde{\beta}_D = 2.26$ ($\tilde{\beta}_R > \tilde{\beta}_M > \tilde{\beta}_D$, O.K.). The deviance on 12 df was about 24, $p \cong 2$ per cent – rather high, but the model is credible. It implies of course that for high concentrations Rotenone is more toxic than Deguelin, and for low concentrations the reverse.

I know nothing about these things, but is it possible? Well, I thought, if Dr Green can wheel in a leading authority to back up his case, why shouldn't I try? So I went and asked Steven Rose, the professor of biology at the Open University. He said, yes, indeed, it is perfectly possible for a poison *A* to be more toxic than poison *B* at high concentrations and less toxic than *B* at low concentrations; he could think of a variety of biochemical mechanisms that would generate such a phenomenon. As I was disappearing through the door he also, I think, murmured something about biologists often being too hung up on linear models – but I may have misheard him . . .

Dr A. J. Lawrance (University of Birmingham): Tonight's paper seems to me to be important, lucid and clever. The section which most interested me was the one dealing with residuals. It is indeed difficult to define widely acceptable residuals or to agree on the properties they should have. Dr Green mentions the importance of uncorrelated residuals for diagnostic determination of model adequacy. Now I must declare my own interest in residuals from non-linear time series. It is a pity that the main ideas in the present paper do not have easier applications to simply dependent data, such as arise from time series. For non-linear autoregressive models the natural definition of residuals is not entirely clear. Peter Lewis and I have taken to defining linear residuals from non-linear models; such residuals are uncorrelated when the model is appropriate, but because of non-linearity, are not independent. The analysis of the higher order dependence properties of uncorrelated residuals is a problem needing further attention. In the time series domain we have worked graphically, and with various correlation functions of the residuals and their squares. Another trouble with most sorts of residuals is that they are omnibus, and not aimed at picking up particular types of departure from the proposed model. With non-linear time series, one may, for instance, consider further sorts of residuals which are sensitive to the directionality of the process, when this is of interest. Methods based on the standard second order properties will overlook this aspect. I also feel that what is needed now in time series is a class of models which can take advantage of the impressive progress made in regression over the last decade.

In conclusion, I join with other contributors in congratulating the author; there is certainly no need to transgress the bounds of RSS english in the discussion of this paper.

The following contributions were received in writing after the meeting.

Mr J. E. Besag (visiting Carnegie-Mellon University): It is a pleasure to congratulate my colleague Peter Green on his wide-ranging paper.

I wish briefly to comment on Section 5.2, concerning resistant variants of least squares and likelihood methods, for it seems there is a danger that the author has added to the confusion he seeks to avoid. Certainly, it may on occasion have been remarked that IRLS updating, as in (27), seeks to minimize $\Sigma\, w(z_i)z_i^2$, a claim which is false unless $w(z) \propto |\, z\, |^{-p}$ and $p < 2$. However, such statements, from which I shall not admit Dr Green once saved me, are entirely incidental, since the minimization of $\Sigma\, w(z_i)z_i^2$, in itself, makes little sense as a general criterion for resistant estimation. The rationale for IRLS here is equivalently, either through the normal equations, as mentioned by Dr Green, or, as I prefer, in terms of successive updating of the weights in weighted least squares (Mosteller and Tukey, 1977, p. 357; Besag, 1981, Section 2). Thus, in the latter framework, one chooses weights, applies these to obtain a weighted least squares fit, calculates corresponding residuals and, in the light of these, forms new weights to apply at the next iteration. Simultaneous minimization is irrelevant: for example, one might even contemplate use of the weight function $w(z) = z^{-2}$, though probably not for long.

Lastly, I remind Dr Green that the suggestion of using IRLS with generalized linear models to obtain resistant variants of maximum likelihood fits is not a new one: see Besag (1981, Section 2).

Professor D. R. Brillinger (University of California, Berkeley): This is a most timely paper. The procedure of iterative reweighting has already found many uses and solved important practical problems, yet its greatest successes would appear to lie ahead. The procedure proves flexible, easy to program, and to have important byproducts. To mention one example of the procedure's power, with a "flick-of-the-wrist" it changes a given estimation method into a robust/resistant one.

The first group of scientists to make substantial use of IRLS were quite possibly the seismologists. They continually find themselves dealing with non-linear models and outlying observations. Jeffreys (1936) made use of the procedure in the construction of tables for the travel times of earthquake waves. Bolt (1960) used it in the determination of the location of earthquakes or explosions given data on the arrival times of the waves at various observatories. Incidentally this last reference introduces the technique of robust/resistant regression some years before statisticians began to study it.

Preisler and I (1983, 1984) have employed iterative reweighting together with numerical integration to fit models involving random effects and latent variables. (Two related references are Bock and Lieberman, 1970 and Hinde 1982.) In the first of our papers we fit a compound Poisson model to multiply-subscripted count data, y_{ijk}, given the covariate x_{ijk}, with y_{ijk} assumed Poisson of mean $\pi + \rho_i x_{ijk} u_{ij}$ given the latent variate u_{ij}. (There is substantial physical background justifying such a detailed model.) GLIM and iterative reweighting lead quickly to the values of the maximum likelihood estimates. In Section 5.1, Green discusses the definition of residuals in non-linear models. There are also difficulties in defining residuals for random effect and latent variable models. In Brillinger and Preisler (1983) we found "uniform residuals" to be useful in detecting an unsuspected phenomenon. (Uniform residuals are defined to be the values one obtains on applying the fitted cumulative distribution functions to the individual observations. They provide a definition of residual for any stochastic model.)

A problem, of quite a different sort, that may be handled via the procedure of iterative reweighting is studied in Brillinger and Preisler (1984). The model $\theta(y_{ij}) = \phi_1(x_{1\,ij}) + \phi_2(x_{2\,ij}) + \epsilon_i + \epsilon_{ij}$ is fit, with the *functions* θ, ϕ_1, ϕ_2 unknown, with ϵ_i a random effect and with ϵ_{ij} error. The solution is obtained by numerical integration, iterative reweighting and repeated use of the ACE procedure of Breiman and Friedman (1984). (ACE determines functions θ, ϕ_1, ϕ_2 to maximize the sample correlation between the values $\theta(y_i)$ and the values $\phi_1(x_{1\,i}) + \phi_2(x_{2\,i})$. It seems destined to find many exciting applications in practical statistics.) The option of using weights was added to ACE as an afterthought. This occurrence, and the importance that Green's paper shows IRLS to have, suggest that developers of statistical algorithms should provide weighted versions whenever possible.

A word of caution needs to be added, and Green does address this issue. The very simplest of non-linear iterations can lead to oscillating and other non-convergent sequences. (May, 1976 is an easily approached reference to this phenomenon.) Now, because of the finite accuracy of computers even linear relationships become effectively non-linear. Proofs of theoretical convergence may not be pertinent. Our experience is that one wants to proceed with the greatest precision allowable.

It seems that IRLS may come to dominate much of statistical computation. Green's paper is a very important one.

Dr R. W. Farebrother (University of Manchester): The author assumes that $\mathbf{A} = B\mathbf{B}^{\mathrm{T}}$ is positive definite but does not consider the possibility of small elements on the diagonal of **B**. In this context Kourouklis and Paige (1981) would recommend the use of

$$\min \mathbf{v}'\mathbf{v} \ \text{ S.T. } \ \mathbf{u} + \mathbf{AD}\boldsymbol{\beta} = \mathbf{AD}\boldsymbol{\beta}^* + \mathbf{Bv}$$

rather than the author's equation (7) which may be rewritten as

$$\min \mathbf{v}'\mathbf{v} \ \text{ S.T. } \ \mathbf{B}^{-1}\mathbf{u} + \mathbf{B}^{\mathrm{T}}\mathbf{D}\boldsymbol{\beta} = \mathbf{B}^{\mathrm{T}}\mathbf{D}\boldsymbol{\beta}^* + \mathbf{v}.$$

A similar problem arises in Section 4.1 when $f(r) \propto \exp(-c \mid r \mid^k)$ as $w(r) = ck \mid r \mid^{k-2}$ and $\psi'(r) = (k-1)w(r)$ will take very large values if r is small and $1 < k < 2$. But the author need not despair of IRLS as Ekblom (1974) and Sposito and others (1977) have published variants of algorithms (17) and (18) which perform satisfactorily in this context and Gentleman's (1974) algorithm is designed for use with non-constant weights.

Further we recommend that the early stages of the iterative process should be subjected to small random shocks as it is possible that the simple iterative procedure will diverge from an apparent solution if subjected to random shocks. See Bohlin's discussion of this point in Wold (1981, pp. 48–50).

Dr M. Green (Polytechnic of North London): My only objection to Dr Green's excellent paper is his suggestion that IRLS is ". . . easily programmed without the aid of packages . . .". His choice of example in 5.3 is interesting since it provides a perfect counter-example to this suggestion. If we follow Finney a little further we find several alterative models that take into account the fact that the "third" poison is a mixture of the other two. I will consider two such alterative models that could be easily understood by a toxicologist lacking the sophistication in Mathematics and Computing necessary to fit such models by programming the IRLS algorithm. Following Finney these models can be called the Independent Action model and the Similar Action model.

Independent Action Model

If we suppose that the poisons act in different physiological ways we could assume that the probability of survival from a mixture is the product of the probabilities of survival from each poison if administered separately. This leads to a model for survival probability of

$$\Phi(\alpha_1 + \beta_1 x_1)\, \Phi(\alpha_2 + \beta_2 x_2),$$

where x_1 and x_2 are log doses for poisons 1 and 2.

Similar Action Model

If we suppose that the poisons act in exactly the same physiological way we could assume that the mixture of poisons at doses d_1 and d_2 is equivalent to a certain dose of poison 1. Taking ρ as the relative potency of poison 1 this leads to a model for kill probability of

$$\Phi(\alpha + \beta \log (d_1 + \rho d_2)).$$

Both models are plausible but face the modeller with the unenviable task of translation into an IRLS algorithm. The overwhelming success of packages such as GLIM is that for many models the user has only to understand the Statistics and how to use the package and facilities are provided for the extension of the class of fittable models relatively easily without recourse to programming an interactive system.

The model for Independent Action can be formulated as a simple case of the Composite Link Function models (Thompson and Baker 1981). The specification and analysis of such models can be automated by use of macros, as described in an unpublished paper by Roger and Colman presented at the GLIM82 conference held at PNL. The model for Similar Action can be analysed simply by use of a macro that calculates the derivatives of the function

$$\alpha + \beta \log (d_1 + \rho d_2)$$

with respect to the parameters α, β and ρ. Use of this macro and a standard GLIM gives estimates for α, β and ρ of -1.834, 1.216 and 0.4940. In both cases GLIM automatically provides the framework for estimation and extra information such as the deviance (36.26 for the Similar Action model) without the user having to program such calculations.

Professor R. I. Jennrich (University of California at L.A.): The author is to be congratulated for identifying a basically simple idea that has many important applications. We have all too few of these. By noting the broad spectrum of applications of IRLS he has illustrated the fundamental role of regression in statistical analysis. Carrying this perhaps to an extreme, I have long been a proponent of the following unified field theory for statistics: "Almost all of statistics is linear regression, and most of what is left over is non-linear regression."

This proposition arose from extensive work on the development of statistical software, primarily BMDP. The heart of almost every program, from ANOVA to multidimensional scaling, looks a good deal like a regression program. Programs that iterate look like non-linear regression, those that do not like linear regression.

Viewing analyses like maximum likelihood and robust estimation from the perspective of IRLS produces not just a convenient computing device, but some basic insights as well. Fitting expected responses to observed responses using IRLS with weights inversely proportional to variances is a very natural thing to do. For the exponential family this is precisely maximum likelihood estimation as has been noted by many including this discussant. As is demonstrated, IRLS provides an insightful way to view robust estimation. The weights play a natural role in reducing the effect of outlying observations and the regression formulation has suggested a number of approximate standard error formulae.

With regard to their use as a computing device, the author notes in his abstract that IRLS algorithms are easily programmed without the aid of packages. I agree, but for me one of the most rewarding uses of IRLS is that of formulating problems so they can be analysed by standard non-linear regression programs without the need to program anything more than the function to be fitted and the weights used. The key to success here as the author points out is finding a formulation that uses ordinary weights rather than generalized weights. The latter are not accepted by most package programs.

I will conclude with a technical comment. The starting point for the discussion of maximum likelihood estimation is the Fisher scoring algorithm

$$I_{\beta\beta}\,\Delta\beta = s_\beta, \tag{$*$}$$

where $s_\beta = dL/d\beta$ is the score vector for the parameter vector β, $I_{\beta\beta} = E_\beta(s_\beta s_\beta^{\mathrm{T}})$ is the corresponding information matrix and $\Delta\beta$ is the Fisher scoring step. In the process of factoring $(*)$ to give his basic equation (3), the author assumes that $L(\beta)$ is a composite function $L(\eta(\beta))$ and that $L(\eta)$ is a likelihood function. The second assumption is not required and dropping it leads to a useful generalization. Assume only that $L(\beta)$ is composite. Let $u = dL/d\eta$ and $A = E_\beta(uu^{\mathrm{T}})$. Then $I_{\beta\beta} = D^{\mathrm{T}}AD$ and $s_\beta = D^{\mathrm{T}}u$ and $(*)$ becomes formally identical to the author's equation (3). All that is lost are two of the three equations above his equation (3). For example, u may not have expectation zero and hence may not be a score vector.

To see what can be gained by this generalization consider the multinomial likelihood $L = \Sigma\, y_i \log p_i$. Let $L(\eta) = \Sigma\, y_i \log \eta_i$ without assuming the η_i sum to 1. Then $u_i = y_i/p_i$, $A_{ij} = \delta_{ij}/p_i$ and $D_{ij} = \partial p_i/\partial\beta_j$. A scoring step is obtained by regressing y on D with weights $w_i = p_i^{-1}$. This is a more symmetric and I think natural formulation of IRLS for the multinomial than that in Section 2.1, and because the weights are now simple any of a variety of standard non-linear regression programs may be used to carry out the required calculations. Moreover, it is not necessary to eliminate a cell to fit the general theory.

Dr A. Pázman (Mathematical Institute, Slovak Academy of Sciences, Bratislava): An alternative to the measures of discrepancy given in Section 5 could be the Rao distance between $\hat{\eta}$ and $\eta(\hat{\beta})$ in the space of the predictors η. By definition, this distance is given by

$$D := \min_\gamma \int_{t_1^\gamma}^{t_2^\gamma} \left[\sum_{ij} \frac{d\gamma_i}{dt}\, E_{\gamma(t)}\left(\frac{\partial L}{\partial \eta_i}\, \frac{\partial L}{\partial \eta_j}\right)\frac{d\gamma_j}{dt}\right]^{\frac{1}{2}} dt \tag{1*}$$

where the minimum is taken over all twice differentiable mappings (curves) from $t \in \langle t_1^\gamma, t_2^\gamma \rangle$ to $\gamma(t) \in R^n$ which are such that $\gamma[t_1^\gamma] = \hat{\eta}$, $\gamma[t_2^\gamma] = \eta(\hat{\beta})$. The distance D is invariant under any differentiable reparameterization of η, and in the normal linear case D^2 is the residual.

There is no difficulty to compute D if $y_1, \ldots, y_n$ are independent and if $L(\eta) = \Sigma_i L_i(\eta_i)$ (as in equation (2.1)). In this case

$$D = \min_{\gamma} \int_{t_1^\gamma}^{t_2^\gamma} \left[\sum_i \left(\frac{d\gamma_i}{dt}\right)^2 E_{\gamma_i(t)} \left[\left(\frac{dL_i(\eta_i)}{d\eta_i}\right)^2 \right]\right]^{\frac{1}{2}} dt.$$

Hence D is simply the Euclidean distance after the reparameterization

$$\eta_i \to \nu_i(\eta_i) := \int \left\{ E_{\eta_i} \left[\left(\frac{dL_i}{d\eta_i}\right)^2 \right] \right\}^{\frac{1}{2}} d\eta_i \quad (i = 1, \ldots, n).$$

For example, if we take the logistic regression in Section 1.1 we have

$$\hat{\eta}_i = y_i / n_i$$

and

$$E_{\eta_i} \left[\left(\frac{dL_i}{d\eta_i}\right)^2 \right] = \frac{n_i}{\eta_i (1 - \eta_i)}.$$

Thus

$$\nu_i(\eta_i) = 2\sqrt{n_i} \arcsin \sqrt{\eta_i}$$

and

$$\begin{aligned} D^2 &= D^2(\hat{\beta}) = \sum_i [\nu_i(\hat{\eta}) - \nu_i(\eta_i(\hat{\beta}))]^2 \\ &= 4 \Sigma_i n_i \left[\arcsin \sqrt{\frac{y_i}{n_i}} - \arcsin \sqrt{\eta_i(\hat{\beta})} \right]^2, \end{aligned} \tag{2*}$$

where $\eta_i(\beta)$ is given in (A), Section 1.1. It follows from (2*) that

$$\begin{aligned} \frac{\partial D^2(\hat{\beta})}{\partial \beta_j} &= -2 \, \Sigma_i n_i \left[\arcsin \sqrt{\frac{y_i}{n_i}} - \arcsin \sqrt{\eta_i(\hat{\beta})} \right] \\ &\quad \times \frac{1}{\sqrt{\{\eta_i(\hat{\beta}) \, [1 - \eta_i(\hat{\beta})]\}}} \frac{\partial \eta_i(\hat{\beta})}{\partial \beta_j} \\ &= \sum_i n_i \frac{y_i/n_i - \eta_i(\hat{\beta})}{[1 - \eta_i(\hat{\beta})] \, \eta_i(\hat{\beta})} + o(\| \hat{\eta} - \eta(\hat{\beta}) \|) \\ &= \frac{\partial L[\eta(\hat{\beta})]}{\partial \beta_j} + o(\| \hat{\eta} - \eta(\hat{\beta}) \|). \end{aligned}$$

Hence for $\hat{\eta}$ which is near to the set $\{\eta(\beta): \beta \in R^m\}$, the ML estimate is almost equal to the estimate obtained by minimizing the Rao distance.

Dr A. N. Pettitt (University of Technology, Loughborough): I was interested to read Dr Green's reference to use of the IRLS method in the linear regression problem in Section 4. I have been working on methods of making estimation techniques robust with censored and grouped observations. One approach is with ranks; see, for example, Pettitt (1983). Another is to use the *EM* algorithm (Dempster *et al.*, 1977) when applied to the regression problem with normal errors and censored and grouped observations. This gives IRLS as the method of estimation. Similarly, as the author notes in Section 4.1, when the *EM* algorithm is applied to the regression problem with "normal/independent" errors, such as the *t* distribution, the IRLS method again results. If one combines the "normal/independent" errors with censored and grouped observations, then, again, the *EM* algorithm results in IRLS estimates.

In this particular case two expectations in the *E* stage have to be carried out and it is important to get the order correct or otherwise IRLS estimates do not result. Explicit results can be found for finite scale mixtures of normal distributions and for the *t* distribution with even degrees of freedom, therefore giving IRLS estimates which are robust to outliers. Estimates can be found using GLIM, for example.

My point, as the author correctly notes, is that when the IRLS estimates result as an application of the *EM* algorithm, then the theory of the *EM* algorithm guarantees that the full iterated values are local maxima or turning points of the likelihood function. Obviously, there is no guarantee of a global maximum unless special circumstances prevail.

No such theory exists for the IRLS method in general. It would appear necessary to compute the observed information, not the expected information, at the fully iterated root, in order to check that the root gives a local maximum and to evaluate the likelihood or quasi-likelihood at the steps of the algorithm in order to check the convergence of the technique.

The IRLS technique may result in convergence, but it should be asked to what; otherwise, statisticians are showing an extremely cavalier attitude to the general problem of optimization.

Dr J. H. Roger (Reading University): The general formulation outlined in this paper includes those models described by Thomson and Baker (1981) as Composite Link functions. Composite Link functions form a technique which allows one to fit any non-linear link function to a model with exponential family error function, using a package such as GLIM, by altering the design matrix at each iteration and is numerically equivalent to that described here as the Newton–Raphson method with Fisher scoring. I have used this approach for fitting both Variety Environment interaction models (Finley and Wilkinson, 1979) and Probits with adjustments *C* and *D* for natural responsiveness (Finney, 1971, p. 126). The major problem encountered is making the iterative process converge when the starting values for the parameters β are far from the optimum. It is necessary to monitor the changes in the log-likelihood at each step and alter the step length appropriately. Such a procedure while certainly not guaranteeing convergence to a global optimum has allowed the solution of several practical problems. The choice for the form for the matrix *A* seems to be of less practical importance. Indeed the use of a suitable parameterization of the model seems to be more crucial.

The package GLIM has been used by several workers to solve certain IRLS problems. However it is not ideal. There is a need for a computing environment which has the array manipulative features of APL, device independent graphics, a powerful macro facility (with editing), text handling and finally a simple IRLS procedure.

Dr A. H. Seheult (University of Durham): It is a pleasure to congratulate my colleague on a lucid and valuable paper.

My comments concern residuals, particularly when the representation (21) does not hold.

Although it may appear natural to define residuals in the observation space, surely this is directly valid only when observations and parameters are structurally related, as in linear models. In the absence of such a structure, a definition of discrepancy in terms of likelihood, such as deviance in equation (20), seems more natural. Often, however, such measures are interpretable as signed squared residuals in the observation space.

In the notation of the paper we would like to write

$$L(\hat{\boldsymbol{\eta}}) - L(\boldsymbol{\eta}(\hat{\boldsymbol{\beta}})) = \Sigma\, \tfrac{1}{2}\, \Delta_i,$$

where $\Delta_i = 2[L_i(\hat{\eta}_i) - L_i(\eta_i(\hat{\boldsymbol{\beta}}))]$ is the deviance associated with y_i; this partition is exact when equation (21) holds. One way of partially resolving the difficulties in the general case referred to by Dr Green is to use jackknife techniques. Denote by $\bar{L}(\boldsymbol{\eta}) = n^{-1} L(\boldsymbol{\eta})$ the average likelihood per observation and define the *pseudo-likelihood* generated by y_i to be

$$L_i^{\circ}(\boldsymbol{\eta}) = n\bar{L}(\boldsymbol{\eta}\,; \mathbf{y}) - (n-1)\,\bar{L}(\boldsymbol{\eta}\,; \mathbf{y}_{(i)}),$$

where $\mathbf{y}_{(i)}$ is $\mathbf{y}$ without y_i. Now define the *pseudo-deviance* associated with y_i to be

$$\Delta_i^\circ = 2(L_i^\circ(\hat{\boldsymbol{\eta}}) - L_i^\circ(\boldsymbol{\eta}(\hat{\boldsymbol{\beta}}))].$$

Note that $\Delta_i^\circ = \Delta_i$ when (21) holds.

Finally, it is interesting to note that $L_i^\circ(\boldsymbol{\eta}) = \log \Pr(y_i \mid \mathbf{y}_{(i)})$ so that $L^\circ(\boldsymbol{\eta}) = \Sigma L_i^\circ(\boldsymbol{\eta})$ is the "log-likelihood" used by Besag (1975, 1977) to determine maximum pseudo-likelihood estimates for conditionally specified spatial models involving awkward normalizing factors in the full likelihood. The above definition of Δ_i° assumes that $\hat{\boldsymbol{\eta}}$ and $\hat{\boldsymbol{\beta}}$ are proper maximum likelihood estimates but this is not necessary and they could be replaced by pseudo estimates.

Mr W. D. Stirling (Massey University, New Zealand): Dr Green's paper extends the IRLS algorithm for generalized linear models to a wide range of other important models. Stirling (1984) independently derived the algorithm for models where responses are independent and involve linear functions of parameters; further applications are described in that paper.

In many problems expected second derivatives cannot be easily specified or evaluated. When $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$, Newton–Raphson iterations are an alternative to Fisher's scoring technique and they can also be found by IRLS. Since

$$\frac{\partial^2 \eta_i}{\partial \boldsymbol{\beta}\boldsymbol{\beta}^{\mathrm{T}}} = \mathbf{0}$$

a step is given by equation (3) with

$$\mathbf{A} = -\frac{\partial^2 L}{\partial \boldsymbol{\eta}\boldsymbol{\eta}^{\mathrm{T}}}.$$

Whereas Fisher's scoring technique ensures that $\mathbf{A}$ is positive semi-definite, making (4) a valid least squares problem, for Newton–Raphson iterations it may not be positive semi-definite. This however is not a problem since (4) still describes the Newton–Raphson step and can also be easily solved. In particular many weighted least-squares algorithms such as Gauss–Jordan, and Givens methods also work with negative weights allowing (4) to be solved when the responses are independent with $\mathbf{A}$ diagonal. Since $\mathbf{D}^{\mathrm{T}}\mathbf{A}\mathbf{D}$ should be positive semi-definite near the solution, and often in a large region of the parameter space, the algorithms are usually still numerically stable. Gill and Murray (1974) suggest a modification to ensure convergence. A consistent estimate of parameter variances is still given by $(\mathbf{D}^{\mathrm{T}}\mathbf{A}\mathbf{D})^{-1}$.

The 2-part iterations for $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\kappa}})$ suggested at the end of Section 3.1 converge very slowly if $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\kappa}}$ are highly correlated. However, if (9) is explicitly soluble for $\boldsymbol{\kappa}$, $\boldsymbol{\kappa}$ can be eliminated by replacing it by its maximum likelihood estimator for fixed $\boldsymbol{\eta}$, $\hat{\boldsymbol{\kappa}}(\boldsymbol{\eta})$ thereby reducing the problem to simple IRLS on $L(\boldsymbol{\eta}, \hat{\boldsymbol{\kappa}}(\boldsymbol{\eta}))$.

Finally I wish to suggest an alternative to the use of *ad hoc* weight functions in Sections 5.2 and 5.3. Robust/resistant M-estimators for normal linear models are equivalent to maximum likelihood assuming a longer-tailed distribution than the normal. Similar methods can be applied to Poisson and binomial data by fitting negative binomial or beta-binomial distributions with the same means. The extra nuisance parameter can be adjusted to give the desired degree of robustness. An advantage of this approach is that the nuisance parameters can also be estimated from the data allowing tests of whether robust/resistant fits are needed.

Dr G. Tunnicliffe Wilson (University of Lancaster): This paper will be extremely valuable to those, like myself, who have been aware for a long time of the importance of IRLS but have failed to appreciate the overall structure of the subject.

My contribution is confined to a small point in Section 4 on Linear Regression. Most of the densities f which one sees in use for the error term are smooth and symmetric, in which case the weight function $w(t)$ is well defined. It is, however, still quite possible to define $w(t)$ for a smooth asymmetric density provided it is centred on its mode.

I have found this useful in a time series context where the residuals in a univariate autoregressive model for a river flow series appeared to have approximately an exponential density. It is pointed out by Dr Green that even for a two-sided exponential density, IRLS cannot be recommended. However, the error model I adopted was the sum of two independent components: $e = e_1 + e_2$, where e_1 is exponential with mean λ and e_2 is Normal $(0, \sigma^2)$. With $\lambda/\sigma \doteq 5$ say, this gives a reasonable model of the observed error density, and has a plausible interpretation. The resulting density is smoothed by the presence of e_2. It is reasonably tractable, its mode is easily found numerically, and the corresponding weight vector w evaluated. Using only one or two steps of IRLS substantially improves the precision of the autoregressive parameter, and the model fit.

Many statistical packages have linear and non-linear least-squares estimation facilities (including time series modelling) with the option of a fixed weight vector supplied by the user. Internally, they solve the equations (17) with fixed weights. When coupled with calculation and loop facilities this allows the full power of IRLS to be exploited. In GENSTAT, for example, calculation of the above-mentioned weight vector required only three extra lines of code. Dr Green's paper stimulated me to try out this example, and I hope it will encourage many others to use the method of IRLS.

Professor C. F. J. Wu (University of Wisconsin, Madison): If the IRLS method is for minimizing an objective function M, be it a log likelihood or not, it is better to use the changes in M for monitoring convergence. This is because the IRLS or a suitable modification yield a monotone sequence of M values, while the behaviour of the sequence of parameter estimates is less predictable. In the context of the *EM* algorithm, Boyles (1983) gave an example that exhibits convergence in the likelihood values but oscillates in the parameter estimates. The convergence behaviours of the *EM* algorithm according to these two criteria are quite different (Wu, 1983). If no such M function exists as in resistant regression, the changes in the left-hand expression of (26) may be used for monitoring convergence. When the iterative estimation of σ is incorporated in Method (I) (17′), more general convergence results are still available from Dempster *et al.* (1980) since their treatment depends on the assumption that the ψ function comes from the normal/independent family. Incidentally this restriction to the *N/I* family can be relaxed by applying the Zangwill's Theorem as in Wu (1983). My only question is on the appropriateness of the definition of Δ_i^* in the dependent situation since it may be quite sensitive to the ordering of the components of η.

I congratulate the author for a stimulating paper which has greatly extended the utility of the IRLS method.

The author replied later, in writing, as follows.

I am very grateful to all the contributors for their stimulating comments. Together they provide a substantial discussion that says much about the power and flexibility of the IRLS method. I have arranged my replies by topic rather than in order of discussant, in the hope of picking out and concentrating on common themes. My intention is to reply to points with a computing emphasis, and then move on to the more statistical aspects, whilst recognizing that the distinction is necessarily blurred. First, however, let me comment on some of the less technical matters raised.

My motivation for looking at the subject of this paper was similar to that of Dr Jϕrgensen. When re-reading Nelder and Wedderburn's paper, I too felt that a more elementary presentation must be possible. The generality arose naturally from this re-formulation and the practical computing details followed from using the method outside generalized linear models, and from attempting to implement resistant alternatives.

Dr Jϕrgensen and Dr Nelder both criticize my "IRLS" terminology. However, the term has the advantage of sounding like a computational method (which it is) and remaining neutral as to the statistical principles leading to it (which is appropriate as there are several). I agree with Dr Nelder that the first syllable of "reweighted" is redundant, but feel that the term has by now caught on rather widely and would be difficult to check. It also makes for a more desirable acronym, especially when generalized.

I am not sure of the utility of the classification of problems requested by Dr Nelder. As I see it, the key feature leading to IRLS is the specification of a regression model as a composite likelihood function $L(\boldsymbol{\eta}(\boldsymbol{\beta}))$. When the IRLS step is reduced to ordinary unweighted least squares, (6) or (7), which is how it will usually be performed, none of the working covariates or the dependent variable remain fixed between iterations, except in extremely special cases. The useful simplifications seem to be essentially those mentioned in Section 1.2: that of a diagonal information matrix $\mathbf{A}$, and that in which there is an underlying linear predictor $\mathbf{X}\boldsymbol{\beta}$ (not necessarily the same as $\boldsymbol{\eta}$) so that the working model matrix changes only in a simple way.

I found the historical comments from Professor Brillinger very interesting; yet again we find that a popular data-analytic technique was fostered outside statistics. With his delightful "unified field theory" and his other remarks, Professor Jennrich has given new insight into the fundamental role of regression in statistical estimation.

As I had expected, a number of contributors to the discussion commented on the use of GLIM for fitting non-standard models by IRLS. Mr Thompson and Mr Burn find the generality of my approach confusing in comparison. It is I think important to retain this generality of description in order to convey the unity of such a wide variety of estimation problems. That is not to imply that in particular problem areas there is no middle ground between this generality and a specific package solution. The user will not necessarily be required to differentiate hundreds of terms. Let me use as an example the multinomial problems with cell probabilities polynomial in the parameters arising in genetics, and mentioned by Mr Thompson, Mr Burn and also Dr Roger. These could be handled directly, but it is more convenient to use the "Poisson trick" and treat the cell frequencies as independent Poisson observations. Thus, $\mathbf{y} \mathrel{\dot\sim}$ Poisson $(\boldsymbol{\eta})$ where $\boldsymbol{\eta} = \mathbf{C}\,\boldsymbol{\gamma}$ and $\boldsymbol{\gamma} = \exp(\mathbf{X}\boldsymbol{\beta})$. In the notation of the paper, $\mathbf{u}$ is $(\mathbf{y}/\boldsymbol{\eta}) - 1$, $\mathbf{A}$ is diag $(1/\boldsymbol{\eta})$, $\mathbf{D}$ is $\mathbf{C}$ diag $(\boldsymbol{\gamma})\mathbf{X}$ and the deviance is $2\,\Sigma\{y_i \log(y_i/\eta_i) - y_i + \eta_i\}$. (Here the division and exponentiation operate component-wise.) The differentiation is already done, yet a package is not needed. This specification is just that given via composite link functions by Thompson and Baker (1981), but expressed in terms of the score function rather than GLIM terminology.

In a similar vein, I am delighted to see that the other Dr Green and I are in such close agreement: I in turn find that *he* provides a perfect counter-example to *his* point about packages. He takes the probit example in Section 5.3 a little further, exploiting the information that one poison is a known (1 : 4) mixture of the others by considering two models suggested by Finney. But the methods he suggests for fitting these two models in GLIM are *different*, and both are different from that by which probit models are fitted without using this mixture information. In fact all of these models, the more general model for synergistic action given by Finney in which the kill probability is $\Phi(\alpha + \beta \log(d_1 + \rho d_2 + \kappa\sqrt{d_1 d_2}))$, the model with complementary-log-log link used by Professor Aitkin, and Professor Lewis's concurrent line probit model, *all* have the same structure: u_i is $((y_i/n_i) - \eta_i)A_{ii}$, where A_{ii} is $n_i/(\eta_i(1-\eta_i))$, and the deviance is

$$2\Sigma\{y_i \log(y_i/n_i\eta_i) + (n_i - y_i)\log((n_i - y_i)/n_i(1-\eta_i))\}.$$

The differences lie only in the functional form of $\boldsymbol{\eta}(\boldsymbol{\beta})$ and the matrix $\mathbf{D}$ of its derivatives. The unity of structure can be reflected in use of essentially the same program, but not, according to Dr Green, if you use GLIM.

Mr Thompson, Professor Aitkin, Mr Burn, Dr Gilchrist and Dr Roger all raise points about estimation problems that are clearly within the scope of IRLS yet are awkward or impossible to handle in GLIM. I agree with all these points, and assume they are directed towards people with more influence in these matters than I. Considerable ingenuity has been shown by some discussants, including Mr Burn, Dr Gilchrist and Dr Tunnicliffe-Wilson, in coding some other difficult problems, including some with non-diagonal $\mathbf{A}$, into GLIM or GENSTAT.

Dr Wilson finds my description of GLIM limited but I think she will find that, apart from her comment about a singular value decomposition for non-diagonal $\mathbf{A}$, her points are contained in the second half of the second paragraph in Section 3.3, admittedly rather briefly.

A number of contributors add some flesh or technical details to the bare bones of the IRLS method as I described it, while others propose alternative procedures. Dr Jϕrgensen is to be congratulated on a very complete coverage of matters such as choice of step length and alternatives to the information matrix $\mathbf{A}$. The basic IRLS method can indeed be adapted to an even larger class of problems, but the programming effort may be greater. As Dr Wilson comments, there is a point at which one gives up and writes a one-off program (or consigns the problem to a general-purpose optimization package). Mr Stirling also discusses choice of A, including possibilities where it does not remain positive semi-definite. These probably take IRLS out of the reach of those who, like Professor Jennrich, prefer to use standard weighted regression packages. The same is true of the very interesting and highly desirable modifications to the IRLS method via constrained optimization, suggested by Dr Farebrother.

Several points are raised about convergence and divergence, and monitoring the iterations. Professor Brillinger reminds us of the unpleasant oscillatory behaviour that may arise in fitting non-linear models and recommends maximum precision, as some of the non-linearity is rounding error. Dr Roger and Professor Wu both advocate monitoring changes in the likelihood. Dr

Roger thereby adjusts the step lengths. The situation described by Professor Wu is disturbing: should there be convergence in the likelihood, but not the parameter estimates, I would like to know! Dr Pettitt notes the absence of any theory in general about the solution to the iteration. The one point on which one can rely is that when iteration ceases at a finite point in parameter space the likelihood equations (1) are satisfied. Quite clearly, many factors may prevent this being an acceptable solution to the estimation problem. My own approach is to note that most of such difficulties will be identified by properly monitoring all aspects of the iteration. It may make for an untidy screen or print-out, but it seems to me essential to write out the current parameter estimates and log-likelihood on every iteration unless the model is an extremely familiar one.

At the end of Section 3.1, I made some brief remarks about use of a two-part iteration in the presence of nuisance parameters. I have found this approach successful in a variety of situations. Dr Gilchrist describes use of this technique when fitting models with parametric link functions, and claims that it is easy to carry this out in GLIM (only for low-dimensional $\boldsymbol{\kappa}$, I would suspect, but this is the usual case). Mr Stirling points out that convergence of this two-part process may be slow if $\boldsymbol{\beta}$ and $\boldsymbol{\kappa}$ are highly correlated, and suggests eliminating $\boldsymbol{\kappa}$ by solving equation (9). To me, this rather begs the question, since the result will generally be to destroy the otherwise simple form for $\boldsymbol{\eta}$ that led me to term the additional parameters a "nuisance" in the first place.

Dr Farebrother recommends incorporating small random disturbances into the early iterations, and this seems to me to be a desirable and convenient complement to the good practice of running the iterative solution from several different starting values: both help to see whether the fitted maximum may be a global optimum. Mr Thompson would like rules for correct parameterization to obtain quick convergence—the experience of Dr Roger shows that these would be valuable—and Mr Ross's mention of stable parameter transformation seems to provide an answer.

Mr Ross also draws our attention to an important rival to IRLS of which I should have been aware. In a model with an additive unimodal log-likelihood, use of the deviance residuals allows the negative log-likelihood to be expressed as a sum-of-squares and hence minimized by the Gauss–Newton method, or some variation of it. Dr Jϕrgensen points out that this can still be regarded as IRLS, with a suitable choice of **A**. In fact, in the notation of the paper, but writing $z_i = \sqrt{\Delta_i}$ (whether signed or not), the Gauss–Newton step is to regress $\mathbf{z}$ on the columns of the matrix $(D_{ij}\, u_i/z_i)$ to obtain $\boldsymbol{\beta}^* - \boldsymbol{\beta}$. Thus, remarkably the programming effort may be even less than with Fisher scoring. The information matrix **A** is still needed for the asymptotic covariance matrix $(\mathbf{D}^{\mathrm{T}}\mathbf{A}\mathbf{D})^{-1}$, or it may be approximated by diag (u_i^2/z_i^2). Can the idea be extended to the case of dependent observations?

Professor Titterington suggests the recursive solution of equation (3). This approach seems particularly attractive for very large data sets, where the inherent slight inefficiency does not matter, and especially where observations are acquired sequentially and processed in real time.

Ten contributors in all refer to the discussion of residuals in Section 5.1. I certainly agree with Dr Jϕrgensen, Mr Thompson and Dr Lawrance that the appropriateness of a definition depends on the use to be made of the residual. In the present context of parametric regression function and probability model, my aim was to suggest methods for assessing the relative importance of the various contributions to the likelihood equations; these might then also be used to determine weights on those contributions in resistant alternatives to maximum likelihood. Viewing the likelihood equations (1) as a sum over components of the predictor vector $\boldsymbol{\eta}$, it is natural to attach residuals to these predictors rather than the observations. Observations are in some contexts somewhat illusory, so that I am not concerned that there is not necessarily a one-to-one correspondence between $\boldsymbol{\eta}$ and **y**. The diagnostic use of residuals will involve examining the form of the probability model $L(\boldsymbol{\eta})$ to find the data points which are associated with a given "discrepant" predictor η_i.

From this standpoint, I agree with the prevailing opinion that the deviances (20) or their signed square roots, the deviance residuals, are appropriate when the likelihood is additive (21). My only purpose in extending the discussion was to treat the case where **A**, the information matrix for $\boldsymbol{\eta}$, is *not* diagonal. Dr Nelder and Professor Wu criticize the sequentially defined Δ_i^* (23). These have the merit of restoring some additivity to the likelihood equations, but the demerit of losing symmetry between the components of $\boldsymbol{\eta}$ by depending on the ordering of these components. Where observations are obtained sequentially, this is perhaps acceptable, but in general it may be preferable to retain the symmetry.

If "greater than" is replaced by "not equal to" in (23) then the $\{\Delta_i^+\}$ reduce to the ordinary deviances $\{\Delta_i\}$ of (20). Mr Thompson and Dr Nelder mention "cross-validatory" residuals, where

Mr Thompson is referring to (20); I would prefer to reserve this term for deviances calculated when the corresponding predictor is omitted from the fitting altogether. Dr Seheult provides an original approach via conditional distributions, again with a cross-validatory flavour, and in which residuals are associated with observations not predictors. But providing that estimates are still obtained by maximizing likelihood not pseudo-likelihood, and that all observations are used in this fitting, Dr Seheult's $\{\Delta_i^\circ\}$ are equal to $\{\Delta_i\}$ in some cases even when (21) fails. For example, this is true for $\mathbf{y} \sim N(\boldsymbol{\eta}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma}$ known, not diagonal, and also for the multinomial distribution with an appropriate interpretation of "missing" observation. How widely does this connection hold? Dr Seheult's suggestion does not succeed in restoring the truth of equation (22), but does provide a useful new viewpoint and deserves further study.

On other residual matters, Mr Davison has provided additional theory for the deviance residuals, which will aid their interpretation, especially for non-Normal linear regression, and Dr Pázman provides an alternative by using the Rao distance; although the interpretation as euclidean distance in a transformed predictor space is appealing, I wonder if these are not too complicated to use in practice? I cannot agree with Dr Jφrgensen that in the simplest non-trivial situation, $\mathbf{y} \sim N(\boldsymbol{\eta}, \Sigma)$ with $\boldsymbol{\Sigma}$ known but not diagonal, the *i*th residual should be $y_i - \eta_i$, scaled. Surely, if for example, a single observation y_i is suspected of being in error, it is its departure from its *conditional* mean given the fitted model and the other observations that should be examined? As above, this leads exactly to Dr Seheult's proposal and the ordinary deviances (20).

Dr Nelder, Dr Gilchrist and Dr Lawrance refer to correlation between residuals. We seem to have learnt to compensate for this in linear models, and I am unhappy at losing the correspondence between residuals and predictors (or observations). Routine use of Gilchrist's adjusted residuals seems entirely practicable and desirable, and the idea should apply to a wider variety of models. Dr Gilchrist's point about assessing $\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}(\hat{\boldsymbol{\beta}})$ directly rather than through $L(\hat{\boldsymbol{\eta}}) - L(\boldsymbol{\eta}(\hat{\boldsymbol{\beta}}))$ is I think answered by considering the case where the former is large and the latter small – does this not imply that apparent discrepancies do not much affect the likelihood for *these* observations and hence are not of interest?

Professor Brillinger's use of uniform residuals is most interesting. It is a pity that statisticians' vision is trained to assess a nominal Normal distribution in residual plots, but of course uniform residuals can always be re-transformed onto any other preferred baseline distribution.

Several contributors mention models and situations where residuals remain difficult to define. I entirely agree with Dr Nelder and Mr Thompson about binary data, and would not advocate these methods for this case. I have nothing to offer Professor Brillinger for latent variable models or Dr Lawrance for non-linear time series, except perhaps that the latter at least might be happy to define deviances sequentially.

Several contributors express puzzlement or concern over the description in Section 5.2 of resistant alternatives to maximum likelihood estimation in non-linear models. Such ideas, at least for generalized linear models, are not new; see Besag (1981) and Pregibon (1982). As to the purpose of such methods, I can only repeat my assertion that it is not the aim of resistant data-analysis (in the context of model-fitting) to provide a unique objective solution, but rather to examine whether any doubt should be cast on a model fitted conventionally. It therefore seems to me eminently reasonable to report a bad fit of, say, a probit model to data from a dosage-mortality experiment, coupled with a considerably better fit to all but a few of the data points.

The "operational difference" between robust and resistant procedures, sought by Professor Atkinson, is surely not eliminated by the fact that at the numerical level the same computations may be carried out in each. In robust regression typically one proceeds with a fair amount of faith in a certain error distribution, but takes out an insurance policy to cover departure from this assumption; the resulting robust procedure is used once, in place of a "classical" method, and the fit is quoted as objective. Resistant methods, on the other hand, may involve the same numerical procedures, but used experimentally at varying levels of resistance, and as a supplement to classical methods.

Mr Thompson's intriguing suggestion that resistant methods fit a "model" that is a mixture of the data and intended model is I think true, even if the idea is a little circular. It ought to aid our understanding of such methods, but I do not know of any principle for fitting the mixing weights in practice that leads to the resistant estimating equations (26) or (28).

Professor Aitkin and Mr Stirling make related points about the desirability of replacing "*ad hoc*" weight functions and tuning constants by prescribed alternative probability distributions and (possibly estimated) additional nuisance parameters. To the extent to which these approaches are actually different, I wonder if we would not then be over-modelling? It should be recalled, incidentally, that since in the resistant approach, the tuning constant used is divided into a *standardized* residual, it does itself have a standardized scale; thus for example with the bi-square function, $B = 9$, 6 and 3 correspond to mild, moderate and high degrees of resistance, for any set of data.

Mr Besag presents an alternative but equivalent rationale for IRLS in resistant methods, and makes the valuable point that there is no need to iterate to convergence, nor indeed any necessity for ultimate convergence at all. This would clearly require careful monitoring of the changing parameter estimates, which as I have stressed is always advisable. There are cases, for example in Table 6, where the fully iterated solution depends on the initial values used, and some where in addition to the fit will not iterate away from the maximum likelihood estimates. There may in general be many solutions to the weighted likelihood equations (28) and some of these will not at all correspond to a "fit" in the ordinary sense. Therefore, we ought perhaps to restrict ourselves to initial estimates that do represent credible models: apart from the exploratory and experimental emphasis, the approach does come close to examining the "robustness" of such models.

Dr Nelder observes that resistant methods are not "resistant" to changes in the link or variance functions. This seems entirely appropriate if we are using a model-based notion of discrepancy. However I certainly agree with Dr Nelder about the risks of using resistant methods with binary data. For example, slight changes in the way in which the method is formulated leads the resistant fit to the vaso-constriction data presented by Pregibon (1982) to go badly wrong: the slopes become infinite and a "perfect" fit is obtained by down-weighting only three observations.

Professor Atkinson notes the omission of any reference in the paper to high leverage and model fit. This is an important matter, but space was restricted. For the binomial/logistic model, I again refer to Pregibon (1982).

Before leaving the subject of resistance, I must return to the probit analysis example that alarms and grieves Professors Aitkin and Lewis. It is of course rather a small data-set (and rather old) to be subjected to such intense scrutiny. I wonder how many different models must be tried before the choice becomes as subjective as the use of resistant methods?

Twelve models or minor variants are suggested in the paper and discussion, including those from Dr Green, and all of these fit considerably better without observations 11, 16 and 17. (If Professor Lewis had estimated his parameters by maximum likelihood, incidentally, he would have found a slightly closer fit; also these estimates are heavily influenced by points 11, 16, 17.)

For a data-set of this size, I would hesitate to abandon an apparently accepted probit transformation. Professor Aitkin's use of the complementary-log-log link yields a reduced deviance overall, but closer inspection of this fit reveals that the almost imperceptible curvature in Deguelin is little altered, while clear curvature and uniformly larger residuals appear for Rotenone.

Further, Professor Lewis finds Rotenone to be more toxic than Deguelin for high concentrations, and less toxic for low: presumably, therefore, the poisons work through different biological mechanisms, so why should the line for the mixture be concurrent with the others?

Among the remaining points with a statistical flavour, I am grateful to a number of contributors for bringing new problems to my attention. These include Dr Pettitt with further examples of the *EM* algorithm reducing to IRLS, namely with censored or grouped data from Normal/independent distributions; Professor Brillinger describes work with random effects and latent variables models, Mr Stirling's paper includes examples with censored data, and negative-binomial and beta-binomial distributions, and Dr Tunnicliffe-Wilson describes experience with asymmetric error distributions in linear regression.

Dr Wilson and Dr Ross are in disagreement over the interpretation of IRLS failing to fit a model. I tend to side with Dr Wilson in asking of what use is an optimal fit of a model that is inappropriate. On the other hand, quite complicated and numerically ill-conditioned models are sometimes appropriate, and we may give up too early. Surely the optimum should still be sought, but the optimization not consigned to a black-box prohibiting visual monitoring of the iteration?

Finally, I am pleased to see that Professor Jennrich shows that any initial formulation is insufficiently general, thus widening even further the range of estimation problems where iteratively reweighted least squares can profitably be applied.

REFERENCES IN THE DISCUSSION

Adena, M. A. and Wilson, S. R. (1982) *Generalized Linear Models in Epidemiological Research: Case Control Studies*. Sydney: Intstat Foundation.

Aitkin, M. (1981) A note on the regression analysis of censored data. *Technometrics*, **23**, 161–163.

——— (1982) Direct likelihood inference. In *GLIM82: Proceedings of the International Conference on Generalized Linear Models* (R. Gilchrist, ed.), pp. 76–86. New York: Springer. (Lecture Notes in Statistics, Vol. 14.)

Aitkin, M. and Clayton, D. (1980) The fitting of exponential, Weibull and extreme value distributions to complex censored survival data using GLIM. *Appl. Statist.*, **29**, 156–163.

Aston, C. E. and Wilson, S. R. (1984) A comment on maximum likelihood estimation in sparse contingency tables. (In the press.)

Atkinson, A. C. (1982) Robust and diagnostic regression analyses. *Commun. Statist.*, **A11**, 2559–2571.

Besag, J. E. (1975) Statistical analysis of non-lattice data. *The Statist'n*, **24**, 179–195.

——— (1977) Efficiency of pseudolikelihood estimation for simple Gaussian fields. *Biometrika*, **64**, 616–618.

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

Bennett, S. (1983) Log-logistic regression models for survival data. *Appl. Statist.*, **32**, 165–171.

Bock, R. D. and Lieberman, M. (1970) Fitting a response model for *n* dichtomously scored items. *Psychometrika*, **35**, 179–197.

Bolt, B. A. (1960) The revision of earthquake epicentres, focal depths and origin times using a high-speed computer. *Geophys. J.*, **3**, 433–440.

Boyles, R. A. (1983) On the convergence of the *EM* algorithm. *J. R. Statist. Soc.* B, **45**, 47–50.

Breiman, L. and Friedman, J. H. (1984) Estimating optimal transformations for multiple regression and correlation. *J. Amer. Statist. Ass.*, to appear.

Brillinger, D. R. and Preisler, H. K.(1984) An exploratory analysis of the Joyner–Boore attenuation data. *Bull. Seismol. Soc. America*, **74** (in press).

Brown, R. L., Durbin, J. and Evans, J. M. (1975) Techniques for testing the constancy of regression relationships over time. *J. R. Statist. Soc.* B, **37**, 149–192.

Cook, R. D. and Weisberg, S. (1983) Comment on Huber (1983). *J. Amer. Statist. Ass.*, **78**, 74–75.

Cox, D. R. and Snell, E. J. (1968) A general definition of residuals. *J. R. Statist. Soc.* B, **30**, 248–275.

Ekblom, H. (1973) Calculation of linear best L_p approximations. *BIT*, **13**, 292–300.

Finlay, K. W. and Wilkinson, G. N. (1963) The analysis of adaption in a plant breeding program. *Aust. J. of Agric. Res.*, **14**, 742–754.

Finney, D. J. (1942) The analysis of toxicity tests on mixtures of poisons. *Ann. Appl. Biol.*, **29**, 82–94.

Frankfurter, M. D. and Jackson, G. (eds) (1929) *The Letters of Sacco and Vanzetti*. London: Constable.

Gilchrist, R. (1981) Calculation of residuals for all GLIM models. *GLIM Newsletter*, **4**, 26–28.

Gill, P. E. and Murray, W. (1974) Newton-type methods for unconstrained and linearly constrained optimization. *Math. Programming*, 7, 311–350.

Goodwin, G. C. and Payne, R. L. (1977) *Dynamic System Identification: Experiment Design and Data Analysis*. New York: Academic Press.

Guerrero, V. M. and Johnson, R. A. (1982) Use of the Box–Cox transformation with binary response models *Biometrika*, **69**, 309–314.

Hinde, J. (1982) Compound Poisson regression models. In *GLIM82: Proceedings of the International Conference on Generalized Linear Models*. (R. Gilchrist, ed.), pp. 109–121. New York: Springer. (Lecture Notes on Statistics, Vol. 14.)

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

——— (1983) Minimax aspects of bounded-influence regression. *J. Amer. Statist. Ass.*, **78**, 66–72.

Jeffreys, H. (1936) On travel times in seismology. *Bur. Centr. Seismol. Travaux Sci.*, **14**, 3–86.

Jørgensen, B. (1983b) The delta algorithm and GLIM. Unpublished manuscript.

Kennedy, W. J., Jr and Gentle, J. E. (1980) *Statistical Computing*, New York: Marcel Dekker.

Kourouklis, S. and Paige, C. C. (1981) A constrained least squares approach to the general Guass–Markov linear model. *J. Amer. Statist. Ass.*, **76**, 620–625.

Krasker, W. S. and Welsch, R. E. (1982) Efficient bounded-influence regression. *J. Amer. Statist. Ass.*, **77**, 595–605.

McCullagh, P. and Nelder, J. A. (1983) *Generalized Linear Models*. London: Chapman and Hall.

Martin, J. T. (1942) The problem of the evaluation of rotenone-containing plants, VI. *Ann. Appl. Biol.*, **29**, 69–81.

May, R. M. (1976) Simple mathematical models with very complicated dynamics. *Nature*, **261**, 459–466.

Mosteller, F. and Tukey, J. W. (1977) *Data Analysis and Regression*. Reading, Mass.: Addison-Wesley.

Pettitt, A. N. (1983) Approximate methods using ranks for regression with censored data. *Biometrika*, **70**, 121–132.

Pregibon, D. (1982b) Score tests in GLIM with applications. In *GLIM82: Proceedings of the International Conference on Generalized Linear Models* (R. Gilchrist, ed.), pp. 87–97. New York: Springer. (Lecture Notes in Statistics, Vol. 14.)

Ross, G. J. S. (1982) Least squares optimization of general log-likelihood functions and estimation of separable linear parameters. In *Compstat 1982* (H. Caussinus *et al.*, eds), pp. 406–411.

Ruppert, D. and Carroll, R. J. (1980) Trimmed least squares estimation in the linear model. *J. Amer. Statist. Ass.*, **75**, 828–838.

Ryder, M. L., Land, R. B. and Ditchburn, R. (1974) Colour inheritance in Soay, Orkney and Shetland sheep. *J. Zool.*, **173**, 477–485.

Scallan, A. (1984) Fitting autoregressive models in GLIM. *GLIM Newsletter* (to appear).

Scallan, A., Gilchrist, R. and Green, M. (1984) Parametric link functions. *Computat. Statist. and Data Anal.*, to appear.

Sposito, V. A., Kennedy, W. L. and Gentle, J. E. (1977) L_p norm fit of a straight line. *Appl. Statist.*, **26**, 114–118.

Stirling, W. D. (1984) Iteratively reweighted least squares for models with a linear part. *Appl. Statist.*, **33**, 7–17.

Titterington, D. M. (1984) Recursive parameter estimation using incomplete data. *J. R. Statist. Soc.* B, **46**, in the press.

Walker, S. H. and Duncan, D. B. (1967) Estimation of the probability of an event as a function of several independent variables. *Biometrika*, **54**, 167–179.

Wold, H. (1981) *The Fix-point Approach to Interdependent Systems*. Amsterdam: North-Holland.

Wu, C. F. J. (1983) On the convergence properties of the *EM* algorithm. *Ann. Statist.*, **11**, 95–103.