ITERATIVELY RE-WEIGHTED LEAST SQUARES AND ITS IMPLEMENTATION IN GLIM4 AND S

G.U.H. Seeber

Anschrift des Autors:

Institut für Statistik
Leopold-Franzens-Universität Innsbruck
Innrain 52
A-6020 Innsbruck
Tel: +43 (512) 507-2744
Fax: +43 (512) 507-2777
E-mail: Gilg.Seeber@uibk.ac.at

SoftStat '93, Heidelberg
Veranstaltungsreihe:

Statistische Algorithmen

Summary

This paper is concerned with a general algorithm for maximum likelihood estimation in a wide class of statistical models, including generalized linear regression and quasi-likelihood. It considers two popular implementations of the algorithm, in GLIM4 and S, and emphasizes on aspects of the user interface and provisions made for fitting more general models than originally intended.

1 Introduction

One of the reasons for the popularity generalized linear models have found is that they have brought much unity into statistical data analysis. Likelihood inference can be performed for a wide class of models of practical importance, which have a common structure, and there is a single algorithm to calculate maximum likelihood estimates. Moreover, there has been available a unique piece of software, GLIM, which provided all the ingredients necessary and useful for statistical modelling, including a simple language to specify models, a common fitting procedure and facilities for diagnosing fitted models. The appearance of a new release, GLIM4, gives the opportunity to review enhancements that have been achieved, particularly in the light of GREEN's (1989) discussion on potential extensions of the class of 'standard' models GLIM3.77 was able to fit.

In this article I do not attempt to give a complete overview on the whole range of applications of the iteratively re-weighted least squares algorithm (IRLS), GLIM's algorithmic heart. Important topics not covered include heteroscedastic linear models (see, e.g., HOOPER, 1993), linear and nonlinear multilevel models (GOLDSTEIN, 1991), bias reduction techniques (FIRTH, 1992a, 1992b). Extensive treatments of the topic may be found in DEL PINO (1989) and GREEN (1984). RUBIN (1983) gives a concise overview, emphasizing robust procedures.

Section 2 introduces the iteratively re-weighted least squares algorithm in the framework of maximum likelihood estimation in curved exponential families. As important instances of applications of IRLS generalized linear models, quasi-likelihood methods and robust procedures are considered in section 3. The final section compares implementations of the algorithm in GLIM4 (FRANCIS, GREEN and PAYNE, 1993) and S (BECKER, CHAMBERS and WILKS, 1988).

2 IRLS for Curved Exponential Families

Consider the log-likelihood function

$$\ell_{\boldsymbol{y}}(\boldsymbol{\vartheta}) = \boldsymbol{y}^t \cdot \boldsymbol{\vartheta} - \boldsymbol{b}(\boldsymbol{\vartheta}) - \boldsymbol{c}(\boldsymbol{y}) \tag{1}$$

for observations $\boldsymbol{y} = (y_1 \; \ldots \; y_n)^t$ of a random variable $\boldsymbol{Y} = (Y_1 \; \ldots \; Y_n)^t$ distributed according to an n-parameter exponential family law, where $\boldsymbol{\vartheta} = (\vartheta_1 \; \ldots \; \vartheta_n)^t \in \Theta$ is the canonical parameter, $\boldsymbol{b}$ is the normalizing function and $\boldsymbol{c}$ relates to the dominating measure. First and second moments are given by

$$\mathrm{E}(\boldsymbol{Y}) = \frac{\partial}{\partial \boldsymbol{\vartheta}} \, \boldsymbol{b}(\boldsymbol{\vartheta}) = \boldsymbol{b}'(\boldsymbol{\vartheta}) \qquad \text{and} \qquad \mathrm{Var}(\boldsymbol{Y}) = \frac{\partial^2}{\partial \boldsymbol{\vartheta} \partial \boldsymbol{\vartheta}^t} \, \boldsymbol{b}(\boldsymbol{\vartheta}) = \boldsymbol{b}''(\boldsymbol{\vartheta}),$$

respectively. As a function of the mean,

$$V(\boldsymbol{\mu}) = b'' \left(b'^{-1}(\boldsymbol{\mu}) \right)$$

is called the variance function. $\boldsymbol{\mu} = \boldsymbol{b}'(\boldsymbol{\vartheta})$ essentially defines a reparameterization. Hence, the log-likelihood function can be expressed, equivalently, in terms of expected values. Score function $\boldsymbol{u}$ and Fisher information $\mathcal{I}$ are therefore given by

$$\begin{aligned} \boldsymbol{u}(\boldsymbol{\mu}) &= \frac{\partial}{\partial \boldsymbol{\mu}} \ell_{\boldsymbol{y}}(\boldsymbol{\mu}) = V(\boldsymbol{\mu})^{-1} \cdot [\boldsymbol{y} - \boldsymbol{\mu}], \\ \mathcal{I}(\boldsymbol{\mu}) &= -\,\mathrm{E}\left[\frac{\partial^2}{\partial \boldsymbol{\mu} \partial \boldsymbol{\mu}^t} \ell_{\boldsymbol{y}}(\boldsymbol{\mu}) \right] = -V(\boldsymbol{\mu})^{-1}. \end{aligned}$$

Using a regression function $\boldsymbol{\eta}$, defined for all $\boldsymbol{\beta} \in \mathbb{R}^k$ and taking values in the *solution locus* $\mathfrak{S} = \{\boldsymbol{\mu} = \boldsymbol{b}'(\boldsymbol{\vartheta}) | \boldsymbol{\vartheta} \in \Theta\}$, and introducing the *systematic component*

$$\boldsymbol{\mu} = \mathrm{E}(\boldsymbol{Y}) = \boldsymbol{\eta}(\boldsymbol{\beta}) \tag{2}$$

defines a *curved exponential family*, see EFRON (1978). To maximize its log-likelihood function *normal equations*

$$\boldsymbol{u}(\boldsymbol{\beta}) = \frac{\partial}{\partial \boldsymbol{\beta}} \ell_{\boldsymbol{y}}(\boldsymbol{\beta}) = \frac{\partial}{\partial \boldsymbol{\beta}} \boldsymbol{\eta}(\boldsymbol{\beta})^t \cdot V(\boldsymbol{\eta}(\boldsymbol{\beta}))^{-1} \cdot \Big[\boldsymbol{y} - \boldsymbol{\eta}(\boldsymbol{\beta}) \Big] = 0. \tag{3}$$

have to be solved for the regression parameters $\boldsymbol{\beta}$. This may be achieved by use of the *scoring algorithm*, which updates a current estimate $\boldsymbol{\beta}^{(k)}$ by

$$\boldsymbol{\beta}^{(k+1)} = \boldsymbol{\beta}^{(k)} + \mathcal{I}(\boldsymbol{\beta}^{(k)})^{-1} \cdot \boldsymbol{u}(\boldsymbol{\beta}^{(k)}).$$

Note that this is a modified Newton-Raphson procedure, where the Hessian is approximated by its expected value.

For the Fisher information of a curved exponential family model we have

$$\mathcal{I}(\boldsymbol{\beta}) = \frac{\partial}{\partial \boldsymbol{\beta}} \boldsymbol{\eta}(\boldsymbol{\beta})^t \cdot V\left(\boldsymbol{\eta}(\boldsymbol{\beta})\right)^{-1} \cdot \frac{\partial}{\partial \boldsymbol{\beta}} \boldsymbol{\eta}(\boldsymbol{\beta}).$$

Starting with an initial guess $\boldsymbol{\beta}^{(0)}$ and writing

$$\boldsymbol{W}^{(k)} = V\left(\boldsymbol{\eta}(\boldsymbol{\beta}^{(k)}) \right)^{-1} \qquad \text{and} \qquad \boldsymbol{U}^{(k)} = \frac{\partial}{\partial \boldsymbol{\beta}} \boldsymbol{\eta}(\boldsymbol{\beta}^{(k)})$$

the $(k+1)$-st iteration of the scoring algorithm is

$$\boldsymbol{\beta}^{(k+1)} = \boldsymbol{\beta}^{(k)} + \left[\boldsymbol{U}^{(k)^t} \cdot \boldsymbol{W}^{(k)} \cdot \boldsymbol{U}^{(k)} \right]^{-1} \cdot \boldsymbol{U}^{(k)^t} \cdot \boldsymbol{W}^{(k)} \cdot \left[\boldsymbol{y} - \boldsymbol{\eta}\left(\boldsymbol{\beta}^{(k)} \right) \right],$$

which may be rearranged as

$$(4) \qquad \boldsymbol{\beta}^{(k+1)} = \left[\boldsymbol{U}^{(k)^t} \cdot \boldsymbol{W}^{(k)} \cdot \boldsymbol{U}^{(k)}\right]^{-1} \cdot \boldsymbol{U}^{(k)^t} \cdot \boldsymbol{W}^{(k)} \cdot \left[\boldsymbol{U}^{(k)} \cdot \boldsymbol{\beta}^{(k)} + \left[\boldsymbol{y} - \boldsymbol{\eta}\left(\boldsymbol{\beta}^{(k)}\right)\right]\right].$$

Note that each step of the algorithm corresponds to the solution of a weighted least squares problem with design matrix

$$\boldsymbol{U}^{(k)} = \frac{\partial}{\partial \boldsymbol{\beta}} \boldsymbol{\eta}(\boldsymbol{\beta}^{(k)}),$$

weight matrix

$$\boldsymbol{W}^{(k)} = V\left(\boldsymbol{\eta}(\boldsymbol{\beta}^{(k)})\right)^{-1}$$

and adjusted dependent variable

$$\frac{\partial}{\partial \boldsymbol{\beta}} \boldsymbol{\eta}(\boldsymbol{\beta}^{(k)}) \cdot \boldsymbol{\beta}^{(k)} + \left[\boldsymbol{y} - \boldsymbol{\eta}\left(\boldsymbol{\beta}^{(k)}\right)\right].$$

That is, for curved exponential families the scoring algorithms is an *iteratively re-weighted least squares* procedure. This suggests a link to geometrical concepts.

The solution locus $\mathfrak{S}$, i.e. the set of expected values in an exponential family model (1), forms a differentiable manifold. Equation (2) restricts the set of 'feasible' expected values to a submanifold $\mathfrak{S}_{\boldsymbol{\eta}}$ of $\mathfrak{S}$. The following geometrical interpretation may be given to the normal equations (3): Residuals $\boldsymbol{y} - \boldsymbol{\eta}(\widehat{\boldsymbol{\beta}})$ are orthogonal, with respect to the inner product defined by $V(\boldsymbol{\eta}(\widehat{\boldsymbol{\beta}}))^{-1}$, to the gradient of the regression function $\boldsymbol{\eta}$ at $\widehat{\boldsymbol{\beta}}$, i.e. the fitted values $\widehat{\boldsymbol{\mu}} = \boldsymbol{\eta}(\widehat{\boldsymbol{\beta}})$ are the *orthogonal projection* of $\boldsymbol{y}$ onto the tangent plane

$$T_{\widehat{\boldsymbol{\mu}}} \mathfrak{S}_{\boldsymbol{\eta}} = \widehat{\boldsymbol{\mu}} + \left\{ \frac{\partial}{\partial \boldsymbol{\beta}} \boldsymbol{\eta}(\widehat{\boldsymbol{\beta}}) \cdot \boldsymbol{\lambda} \,\middle|\, \boldsymbol{\lambda} \in \mathbb{R}^k \right\}.$$

Thus, for curved exponential families the optimization problem of maximizing the log-likelihood is equivalent to the geometrical problem of an orthogonal projection. As a weighted least squares problem each cycle of the scoring procedure can be regarded as an approximation to this. Many diagnostic techniques well established in normal linear regression theory are therefore easily generalized for use in the curved exponential family framework.

Note that no assumptions have been made about the structure of the regression function $\boldsymbol{\eta}$. It may contain parameters other than $\boldsymbol{\beta}$, possibly parametric as well as non-parametric components; see, e.g., Green (1984) and Hastie and Tibshirani (1990). In this case, IRLS can be part of a more general algorithm for both smoothing and maximization of the likelihood.

(Almost) all we need is $\boldsymbol{\eta}$ to be differentiable. To have IRLS implemented in full generality, facilities for symbolic differentiation — similar to the S-interface to `Mathematica`, see Cabrera and Wilks (1993) — would certainly be of advantage.

It may be worthwhile to remark that Jørgensen (1984) suggests a modification of the scoring (also, the Newton-Raphson) procedure, which he calls the delta algorithm, where the weight matrix in (4) is to be replaced by some suitably chosen symmetric and positive definite matrix $\boldsymbol{K}(\boldsymbol{\mu})$ and $\ell_{\boldsymbol{y}}$ may be parametrized via an 'auxiliary parameter' $\boldsymbol{\mu}$ other than the mean.

3 Other Instances of IRLS

This section briefly describes some of the major applications of the iteratively re-weighted least squares algorithm to classes of models or methods that have found much attention in statistical data analysis.

3.1 Exponential Dispersion and Generalized Linear Models

Introducing a dispersion parameter σ^2 into an exponential family log-likelihood (1),

$$\ell_{\boldsymbol{y}}(\boldsymbol{\vartheta}, \sigma^2) = \frac{\boldsymbol{y}^t \cdot \boldsymbol{\vartheta} - \boldsymbol{b}(\boldsymbol{\vartheta})}{\sigma^2} - \boldsymbol{c}(\boldsymbol{y}, \sigma^2),$$

does not mean any essential change in the normal equations (3) for the regression parameters $\boldsymbol{\beta}$ of a curved exponential family model, provided the maximum likelihood estimate for σ^2 exists. The IRLS procedure for *exponential dispersion models* is therefore the same as given in (4).

JØRGENSEN (1987) provides a discussion and examples of exponential dispersion models.

As a special case *generalized linear models* (c.f. MCCULLAGH and NELDER (1989)) have mutually independent, one parameter exponential family components and a simple structure of the regression function $\boldsymbol{\eta}(\boldsymbol{\beta})$, where the mean μ_i relates to a linear predictor via a link function g

$$g(\mathrm{E}(Y_i)) = g(\mu_i) = \sum_{j=1}^{k} x_{ij} \cdot \beta_j = x_{i\cdot} \cdot \boldsymbol{\beta},$$

or, vice versa,

$$\mathrm{E}(Y_i) = \mu_i = h(x_{i\cdot} \cdot \boldsymbol{\beta}) = \boldsymbol{\eta}(\boldsymbol{\beta}).$$

Using the notation $\boldsymbol{b}(\boldsymbol{\vartheta}) = (b(\vartheta_1) \ \dots \ b(\vartheta_n))^t$, $\boldsymbol{g}(\boldsymbol{\mu}) = (g(\mu_1) \ \dots \ g(\mu_n))^t$ and $\boldsymbol{D}(\boldsymbol{\beta}) = \mathrm{diag}\{h'(x_{1\cdot} \cdot \boldsymbol{\beta}), \dots, h'(x_{n\cdot} \cdot \boldsymbol{\beta})\}$ we have

$$\frac{\partial}{\partial \boldsymbol{\beta}} \boldsymbol{\eta}(\boldsymbol{\beta}) = \boldsymbol{D}(\boldsymbol{\beta}) \cdot \boldsymbol{X},$$
$$V(\boldsymbol{\eta}(\boldsymbol{\beta})) = \mathrm{diag}\{\boldsymbol{b}''(x_{1\cdot} \cdot \boldsymbol{\beta}), \dots, \boldsymbol{b}''(x_{n\cdot} \cdot \boldsymbol{\beta})\}.$$

With $\boldsymbol{D}^{(k)} = \boldsymbol{D}(\boldsymbol{\beta}^{(k)})$ the $(k+1)$-st iteration of the scoring algorithm becomes

$$\boldsymbol{\beta}^{(k+1)} = \left[\boldsymbol{X}^t \cdot \boldsymbol{D}^{(k)} \cdot \boldsymbol{W}^{(k)} \cdot \boldsymbol{D}^{(k)} \cdot \boldsymbol{X}\right]^{-1} \cdot \boldsymbol{X}^t \cdot \boldsymbol{D}^{(k)} \cdot \boldsymbol{W}^{(k)} \cdot \boldsymbol{D}^{(k)} \cdot \left[\boldsymbol{X}\boldsymbol{\beta}^{(k)} + \frac{\partial}{\partial \boldsymbol{\mu}} \boldsymbol{g} \left(\boldsymbol{y} - \boldsymbol{\eta}(\boldsymbol{\beta}^{(k)})\right)\right].$$

Note that there is no need for an initial guess $\boldsymbol{\beta}^{(0)}$, $\boldsymbol{g}(\boldsymbol{y})$ may serve as a starting value $\boldsymbol{X} \cdot \boldsymbol{\beta}^{(0)}$. For *canonical link functions*, i.e. $\boldsymbol{g} = \boldsymbol{b}'^{-1}$ or $\vartheta_i = x_{i\cdot} \cdot \boldsymbol{\beta}$, this further simplifys to

$$\boldsymbol{\beta}^{(k+1)} = \left[\boldsymbol{X}^t \cdot \boldsymbol{W}^{(k)} \cdot \boldsymbol{X}\right]^{-1} \cdot \boldsymbol{X}^t \cdot \boldsymbol{W}^{(k)} \cdot \left[\boldsymbol{X} \cdot \boldsymbol{\beta}^{(k)} + \boldsymbol{W}^{(k)} \cdot \left(\boldsymbol{y} - \boldsymbol{\eta}(\boldsymbol{\beta}^{(k)})\right)\right],$$

because, in this case, $\boldsymbol{D}(\boldsymbol{\beta}) = V(\boldsymbol{\eta}(\boldsymbol{\beta})) \cdot \boldsymbol{X}$.

For each instance the scoring algorithm is an iteratively re-weighted least squares algorithm. Inner products and tangent planes projected onto are different, however.

There are several extensions to generalized linear models for which maximum likelihood estimates may be obtained by IRLS, including GLMs with *composite link functions* (THOMPSON and BAKER, 1981) $\boldsymbol{\eta}(\boldsymbol{\beta}) = \boldsymbol{h}\left(\boldsymbol{\alpha}^t \cdot \boldsymbol{C} \cdot (h(x_{1\cdot} \cdot \boldsymbol{\beta}) \ \ldots \ h(x_{n\cdot} \cdot \boldsymbol{\beta}))^t\right)$ or *parametric link functions* (SCALLAN, GILCHRIST and GREEN, 1984) $\boldsymbol{\eta}(\boldsymbol{\beta}) = (h(x_{1\cdot} \cdot \boldsymbol{\beta}, \alpha) \ \ldots \ h(x_{n\cdot} \cdot \boldsymbol{\beta}, \alpha))^t$.

3.2 Quasi-likelihood Models and Estimating Equations

The normal equations (3) for curved exponential families only depend on the regression function $\boldsymbol{\eta}$, specifying first moments, and the variance function V. If a full likelihood function is not available or intractable inference can be based on the *quasi-score function*

$$\boldsymbol{u}(\boldsymbol{\beta}) = \frac{\partial}{\partial \boldsymbol{\beta}} \boldsymbol{\eta}(\boldsymbol{\beta})^t \cdot V(\boldsymbol{\eta}(\boldsymbol{\beta}))^{-1} \cdot \Big[\boldsymbol{y} - \boldsymbol{\eta}(\boldsymbol{\beta})\Big]$$

instead of the derivatives of $\ell_{\boldsymbol{y}}$. If there exists a function $Q_{\boldsymbol{y}}(\boldsymbol{\beta})$ such that its derivatives are equal to $\boldsymbol{u}(\boldsymbol{\beta})$ it is called a *quasi-likelihood function.* Quasi-likelihood functions share many of the properties of ordinary log-likelihood functions. In particular, solving the *estimating equation* $\boldsymbol{u}(\boldsymbol{\beta}) = 0$ corresponds to performing maximum-likelihood estimation for fully specified models. See MCCULLAGH (1991) for an overview and examples.

It should be noted that, in general, a quasi-likelihood function need not exist. In that case the theory of estimating equations – see, e.g., GODAMBE and HEYDE (1987) – provides the justification for this method. Clearly, IRLS can be used for solving the estimating equation. Its interpretation as a sequence of approximations to an orthogonal projection remains valid whether or not a quasi-likelihood function does exist.

3.3 Models with a Linear Part, Robust Procedures

IRLS arises naturally in the framework of exponential families, there are, however, important applications with different context. STIRLING (1984), e.g., considers log-likelihood functions of the form

$$\ell_{\boldsymbol{y}}(\boldsymbol{\beta}) = \sum_{i=1}^{n} f(y_i | x_{i\cdot} \cdot \boldsymbol{\beta}), \tag{5}$$

which depend on the parameters $\boldsymbol{\beta}$ only through the linear predictor $x_{i\cdot} \cdot \boldsymbol{\beta}$. For such a *model with a linear part* the Newton-Raphson procedure reduces to an iteratively re-weighted least squares algorithm with weight matrix

$$W^{(k)} = \operatorname{diag}\{-f''(y_1|x_{i\cdot} \cdot \boldsymbol{\beta}^{(k)}), \ldots, -f''(y_n|x_{i\cdot} \cdot \boldsymbol{\beta}^{(k)})\}$$

and adjusted dependent variable $\boldsymbol{z}^{(k)}$ with

$$z_i^{(k)} = x_{i\cdot} \cdot \boldsymbol{\beta}(k) - \frac{f'(y_i|x_{i\cdot} \cdot \boldsymbol{\beta}^{(k)})}{f''(y_i|x_{i\cdot} \cdot \boldsymbol{\beta}^{(k)})}.$$

Examples for the method's use include models involving the negative binomial or beta-binomial distributions, and models fitted to grouped and censored data.

For linear models some M-estimators are maximum likelihood estimators under special distributional assumptions, e.g. symmetric distributions that have thicker tails than the normal to accomodate for extreme observations. These distributions cannot be found in exponential families, but models usually have the structure of the log-likelihood (5). In this case, IRLS is an EM-algorithm (Dempster, Laird and Rubin, 1977) and general convergence results about EM apply to IRLS, for which is a lack in the general case.

4 Software Implementations

Both Glim and S appeal to applied statisticians, who are in need of a software package that offers all the tools useful and necessary for practical data analysis while maintaining flexibility and supporting programming facilities. S, adopting an object-oriented approach, is no doubt the 'modern' solution, 'old fashioned' Glim still has, also no doubt, its virtues and continues to be one of the major packages, particularly with the new realease, which offers many enhancements, including high quality graphics and extensions to the model formulae syntax.

Having experience with both of the packages I decide on which to use upon the particular problem at hand. It may also be just a matter of taste which to prefer in many circumstances. In this section only the implementation of IRLS is under consideration. Some familiarity with Glim and S will be assumed.

4.1 Glim4

To quote Francis, Green and Payne (1993) 'Model fitting in Glim4 has been entirely rewritten. The model formulae facilities have been extended; the syntax and internal code has been rationalised and the range of options substantially extended. Error checking has been substantially enhanced, and model fitting is now more accurate.'

In Glim3.77 during each cycle (i.e. iteration of IRLS) of the fit (invoked through a `$fit` *model formula* directive) the weighted least squares problem was numerically solved by a Gauss-Jordan procedure. For the new version the Givens algorithm, which is considered much more accurate, has been implemented as a default. In what follows I will assume that Gauss-Jordan has been selected by use of the `$method Jordan` directive. As a consequence the SSP matrix is being formed explicitly und subsequently inverted during each cycle.

As with previous releases `own` models, i.e. generalized linear models with 'non-standard' error-link combinations, can be specified by the user. Now, this is done more naturally by use of the `$error` and `$link` directives. To set things up, in particular to compute starting values, a macro, whose name has been given as an argument to the `$initial` directive, is invoked once in each fit.

As a completely new feature the user may interfere with the fitting procedure at various stages of each cycle in order to modify IRLS components, thus declaring an `own` algorithm. To modify system structures the following directives may be used:

- `$initial` *macro*: Called once in each fit, at least starting values have to be set by *macro*.

- `$load` *macro*: *macro* may be used to modify the 'working triangle', i.e. the lower triangle of the SSP matrix, before it has been inverted.
- `$load` *vector*: During each cycle the values held in *vector* will be added to the diagonal of the working triangle before inverting.
- `$method *` *macro*: To calculate the regression function $\boldsymbol{\eta}$ and its derivatives *macro* will be called.
- `$method * *` *macro*: Iterative weights and the working variate, i.e. the adjusted dependent variable, may be modified by *macro*.

4.2 S

CHAMBERS and HASTIE (1992) claim that 'S is a natural environment for exploring generalized linear models and for creating suitable data structures for their representation and fitting. For example, the link and variance functions are packaged up in the *family* argument to the `glm` function, and are themselves S functions; typically they are very simple, and as such they can be easily modified and new link and variance functions can be created.'

Among others `glm` takes the `family=`*family* argument as a parameter. *family* is a generator function that evaluates to a family object. Components of family objects include seven S functions to compute the link function, its inverse and derivative, to calculate the variance function, the deviance and the weights, and to do the initialization for IRLS. There are predefined families for standard GLMs. Moreover, functions such as `quasi`, `make.family` or `robust` are supplied to modify or build families, always assuming independent observations.

IRLS actually is performed by the function `glm.fit`, which assumes that the regression function is of GLM type (i.e. means are related to a linear predictor via a link function) and observations are mutually independent. However, `glm.fit` can be replaced by a user written procedure.

4.3 Discussion

At first sight many of the proposals in GREEN (1989) have been realized in GLIM4. The basic building blocks of the iteratively re-weighted least squares algorithm are accessible to the user and can be modified to the user's choice. This can be quite cumbersome, however, because computations tend to be slow with `own` models and algorithms, and, even more important, the GLIM4 language lacks of several convenient data structures and mathematical functions. There is now available an `array` similar to Fortran arrays, but it cannot be used as a substitute for matrix as a basic data structure, since arithmetic can only be performed on its representation as a column vector.

Compared to GLIM4 there is some richness in the S language. But as an interpreted language it also slows down numerical calculations. IRLS computations are coded in Fortran and the basic elements of the algorithm are not accessible to the user. With respect to the user interface to the numerical algorithm GLIM4 seems to do better.

It should be noted that on many platforms Fortran and/or C code can be linked statically or dynamically to either of the packages. In the case of S a few such programs are available on `statlib` for anonymous ftp (on `lib.stat.cmu.edu`), where scripts for S and macros for GLIM may be found as well.

There are two discussion lists, `s-news` and `glim-l`, serving as an electronic forum for users and developers of the respective packages. The former is very active.

References

BECKER, R.A., CHAMBERS, J.M. and WILKS, A.R. (1988). *The New S Language. A Programming Environment for Data Analysis and Graphics*, Pacific Grove, Cal.: Wadsworth.

CABRERA, J.F. and WILKS, A.R. (1993). An interface from S to `Mathematica`. AT&T Bell Labs Statistics Research Report.

CHAMBERS, J.M. and HASTIE, T.J. (eds.) (1992). *Statistical Models in S*, Pacific Grove, Cal.: Wadsworth.

DEL PINO, G. (1989). The unifying role of iterative generalized least squares in statistical algorithms (with comments and rejoinder). *Statistical Science* **4**, 394–408.

DEMPSTER, A.P., LAIRD, N.M. and RUBIN, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society*, Ser. B **39**, 1–38.

EFRON, B. (1978). The geometry of exponential families. *Annals of Statistics* **6**, 362–376.

FIRTH, D. (1992a). Bias reduction, the Jeffreys prior and GLIM. *Advances in* GLIM *and Statistical Modelling* (L. Fahrmeir, B. Francis, R. Gilchrist and G. Tutz, eds.), 91–100. New York: Springer.

_____ (1992b). Generalized linear models and Jeffreys prior: an iterative weighted least squares approach. *Computational Statistics* (Y. Dodge and J. Whittaker, eds.) 553–557. Heidelberg: Physica Verlag.

FRANCIS, B., GREEN, M. and PAYNE, C. (1993). *The GLIM4 Manual.* Oxford: Oxford University Press.

GODAMBE, V.P. and HEYDE, C.C. (1987). Quasi-likelihood and optimal estimation. *International Statistical Review* **55**, 231–244.

GOLDSTEIN, H. (1991). Nonlinear multilevel models, with an application to discrete response data. *Biometrika* **78**, 45–51.

GREEN, P. (1984). Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives (with discussion). *Journal of the Royal Statistical Society*, Series B **46**, 149–192.

_____ (1989). Generalised linear models and some extensions. Geometry and algorithms. *Glim 89 and Statistical Modelling* (A. Decarli, B. Francis, R. Gilchrist and G.U.H. Seeber, eds.), 26–36. New York: Springer.

HASTIE, T.J. and TIBSHIRANI, R.J. (1990). *Generalized Additive Models.* London: Chapman and Hall.

HOOPER, P.M. (1993). Iterative weighted least squares estimation in heteroscedastic linear models. *Journal of the American Statistical Association* **88**, 179–184.

JØRGENSEN, B. (1984). The delta alorithm and GLIM. *International Statistical Review* **52**, 283–300.

_____ (1987). Exponential dispersion models (with discussion). *Journal of the Royal Statistical Society* Series B **49**, 127–162.

MCCULLAGH, P. (1991). Quasi-likelihood and estimating functions. *Statistical Theory and Modelling* (D.V. Hinkley, N. Reid and E.J. Snell, eds.), 263–286. London: Chapman and Hall.

_____ and NELDER, J.A. (1989). *Generalized Linear Model. Second Edition.* London: Chapman and Hall.

RUBIN, D.B. (1983). Interatively reweighted least squares. *Encyclopedia of Statistical Science* (N.L. Johnson and S. Kotz, eds.), vol. 4, 272–275. New York: Wiley.

SCALLAN, A., GILCHRIST, R. and GREEN, M. (1984). Fitting parametric link functions in generalised linear models. *Computational Statistics & Data Analysis* **2**, 37–49.

STIRLING, W.D. (1984). Iteratively reweighted least squares for models with a linear part. *Applied Statistics* **33**, 7–17.

THOMPSON, R. and BAKER, J.R. (1981). Composite link functions in generalized linear models. *Applied Statistics* **30**, 125–131.