INFKEEP, a Genstat procedure for calculating influence functions and associated standard errors.

Roger Newson

AFRC Arable Crops Research, Rothamsted Experimental Station, Harpenden, Herts, UK

September 6th, 1994

1. Introduction

The method of M-estimators is a method of parameter estimation, derived by generalising maximum likelihood theory to the case where the deviance function being minimised is not necessarily related to the log likelihood in any simple way. The theory is given briefly in Ruppert (1985) [1], and more rigorously in Chapter 7 of Serfling (1980) [2]. In general, we say that we are minimising the mean of a function $\rho(X,\theta)$ with respect to θ given X, where X is a random variable and θ is a parameter. The first derivative of $\rho(X,\theta)$ with respect to θ is denoted $\psi(X,\theta)$, and its mean is assumed to be zero when the mean of $\rho(X,\theta)$ is minimised. The corresponding second derivative is denoted $\psi'(X,\theta)$. If θ is a vector of length p, say, then $\psi(X,\theta)$ is also a p-vector and $\psi'(X,\theta)$ is a $p \times p$ matrix. Given a vector of n data units $X_1,\ldots X_n$, usually assumed independent and identically distributed with distribution function $F(\cdot)$ (which may also be multivariate), then the M-estimator is the value $\hat{\theta}$ satisfying

$$\sum_{j=1}^{n} \psi(X_j,\hat{\theta}) = 0. \tag{1.1}$$

Under "reasonable" assumptions, $\hat{\theta}$ is a consistent estimator of θ_F, the value satisfying

$$E_F\left[\psi(X,\theta_F)\right] = 0, \tag{1.2}$$

where $E_F[\cdot]$ denotes expectation assuming that the distribution function of X is $F(\cdot)$.

To derive the distribution of $\hat{\theta}$ around θ_F, we use the influence curve, defined as

$$IC(X,\theta) = -\left\{E_F\left[\psi'(X,\theta)\right]\right\}^{-1}\psi(X,\theta). \tag{1.3}$$

The influence curve is, in general, a p-vector of random variables for each value of θ. If the

true distribution function is $F(\cdot)$, then in the limit, as $n \to \infty$, the distribution of $n^{\frac{1}{2}}(\hat{\theta}-\theta_F)$ tends to a multivariate normal, with dispersion matrix equal to that of $IC(X,\theta_F)$.

Note that, in the special case where $F(\cdot)$ belongs to a family of distributions parameterised by θ, and $\rho(X,\theta)$ is $-2 \times$ log likelihood of θ given X, then we have the case of classical maximum likelihood. In this case, we can assume (see Section 2.11 of Silvey, 1975 [3]) that

$$D_F\left[\psi(X,\theta_F)\right] = E_F\left[\psi'(X,\theta_F)\right], \tag{1.4}$$

where $D_F[\cdot]$ denotes dispersion matrix, assuming that $F(\cdot)$ is the distribution function of X. It follows that the dispersion matrix of the influence function is

$$\begin{aligned} D_F\left[IC(X,\theta_F)\right] =& E_F\left[\psi'(X,\theta_F)\right]^{-1} D_F\left[\psi(X,\theta_F)\right] E_F\left[\psi'(X,\theta_F)\right]^{-1} \\ =& E_F\left[\psi'(X,\theta_F)\right]^{-1}. \end{aligned} \tag{1.5}$$

In the more general case, we wish to derive confidence limits for θ_F around the point estimate $\hat{\theta}$ without assuming the equality (1.4). Usually, the dispersion matrix of the true influence curve can be estimated consistently by the sample dispersion matrix of the sample influence curve, whose j'th element is defined as

$$\widehat{IC}(X_j) = -\left\{n^{-1}\sum_{k=1}^{n}\psi'(X_k,\hat{\theta})\right\}^{-1}\psi(X_j,\hat{\theta}). \tag{1.6}$$

The motivation for using the sample influence curve, instead of estimating the information matrix and using the equality (1.4), is that a model can be useful without being perfect in the sense of fully describing the dispersion of the ψ-function. In particular, the variable "X" may be a vector of a "y-variable" and "x-variables", and the model may be good at regressing the y-variable with respect to the x-variables, but may be inaccurate in that it does not allow for heteroscedasticity, overdispersion and underdispersion. These three inaccuracies may be unimportant in choosing the best estimates of the regression parameters, but may be important in estimating standard errors. Moreover, some ρ-functions are chosen not for being exactly proportional to the log likelihood under a particular model, but for the fact

that the estimates arising from them are robust to deviations from the model, such as a small subpopulation of outliers. (See Huber, 1981 [4].)

Influence functions are also used as diagnostic indicators, to indicate the influence of a particular observational unit on an estimated parameter. For these uses, see Reid, 1983 [5], and references cited there.

The present report describes a comprehensive Genstat procedure INFKEEP, which takes, as input, a regression save structure produced by fitting the parameters of a general linear or nonlinear model, and gives, as output, the sample influence curve, and the standard errors for the parameters derived from it. The procedure makes it possible to use the excellent optimisation software of Genstat for confidence interval estimation using nonstandard ρ-functions, or standard ρ-functions that are imperfect because of heteroscedasticity, overdispersion or underdispersion.

The present author has used INFKEEP extensively for fitting general linear and nonlinear models to overdispersed insect trap counts. The standard errors yielded by INFKEEP have typically been larger than those estimated by standard maximum-likelihood theory assuming a Poisson error distribution, sometimes by a factor of 5, although factors of 2 or less are more common.

The next two Sections describe the methods used by INFKEEP in the respective cases of nonlinear models and general linear models. There follows a final Section, suggesting possible amendments to INFKEEP.

2. Nonlinear models

This case is the simpler of the two, and is therefore given first. In general, for general linear or nonlinear models, Genstat minimises a deviance, which is a sample sum of ρ-functions of form

$$\sum_{i=1}^{n} \rho(y_i, \theta), \tag{2.1}$$

where n is the sample number, y_i is the i'th value of the y-variate, and θ is the parameter (usually vector). The form of $\rho(y_i, \theta)$ is decided by the DISTRIBUTION option of the MODEL statement, and can be one of 5 forms, each given as a function of the observed y_i and the fitted values f_i which depend on the parameter θ. These forms are as follows:

Normal	$(y_i - f_i)^2$
Poisson	$2[y_i \log(y_i/f_i) - (y_i - f_i)]$
Binomial	$2\{y_i \log(y_i/f_i) + (n - y_i) \log[(n - y_i)/(n - f_i)]\}$
Gamma	$2[(y_i - f_i)/f_i - \log(y_i/f_i)]$
Inverse Normal	$[(y_i - f_i)^2/(y_i f_i^2)]$

(2.2)

(See Section 8.5 of the Genstat manual [6].) If we wish to use an arbitrary nonnegative ρ-function not covered in (2.2), then we can simply calculate its square root as the f_i and fit it to a y-variate of zeros, using the Normal option. Either way, the dependency of $\rho(y_i, \theta)$ on θ is mediated solely by f_i, so the ψ-function is given by

$$\psi(y_i, \theta) = \frac{\partial f_i}{\partial \theta} \frac{\partial}{\partial f_i} \rho(y_i, \theta). \tag{2.3}$$

The first factor of the right-hand side of (2.3) is given as the GRADIENTS output parameter of the directive RKEEP. (See Sections 8.1 and 8.6.4 of the Genstat manual [6]. It is a particular selling point of Genstat that it calculates these gradients by numerical differentiation, saving the users from having to define their own derivatives.) The second factor is calculated by INFKEEP using an expression chosen after consulting the DISTRIBUTION

option of RKEEP, and the two factors are multiplied to evaluate the PSI output parameter of INFKEEP.

Given PSI, it remains to calculate the influence curve using Equation (1.6). To do this, INFKEEP uses the INVERSE output parameter of RKEEP, which, in the case of a nonlinear model, is an estimate of the inverse of the matrix of second derivatives of half the deviance with respect to the parameters. INFKEEP then multiplies INVERSE by $n/2$ to evaluate its output parameter INVPSIPRIME, equal to the first factor in the right hand side of (1.6).

If the RCYCLE command has been used and the METHOD option set to NewtonRaphson, then Genstat calculates INVERSE using numerical second-order differentiation. If METHOD is GaussNewton, then, apparently, INVERSE is calculated using only the first-order numerical differentiation that produced GRADIENTS, followed by calculating a sum of squares and products. The validity of this method depends on the identity (1.4), and the point of INFKEEP is that it does not require this assumption. In the case of a nonlinear model, INFKEEP therefore checks the rsave structure to ensure that the Newton-Raphson method has been used, and, if not, it EXITs, explaining why. If the user wishes to fit the model using another method (to save CPU time), then this fitting should be followed by a single iteration using Newton-Raphson, to set INVERSE to the correct value.

Having calculated PSI and INVPSIPRIME, INFKEEP can then evaluate the influence curves in its output parameter INFCUR, using Equation (1.6). The sample dispersion matrix of the variates in INFCUR is then calculated, and divided by n to produce the estimated parameter variance-covariance matrix in the output parameter VCOV. The parameters CORR and SE are then calculated from VCOV, using CORRMAT and the SQRT of the diagonal, respectively.

3. General linear models

The methods used for general linear models (GLMs) are the same as those for nonlinear models, except for some differences in the calculation of the parameters GRADIENTS and INVERSE. In this Section, the notation will be slightly different, for consistency with that usually used when talking about GLMs ,as in Section 8.5 of the Genstat manual [6] or Chapter 2 of McCullagh and Nelder, 1983 [7]. The vector of parameters will be denoted β instead of θ, and the ρ-, ψ- and ψ'-functions will be denoted $\rho(Y, X, \beta)$, $\psi(Y, X, \beta)$ and $\psi'(Y, X, \beta)$, respectively, where X is a p-vector of x-values related to the "fitted" value f by a relation of the form

$$X^T\beta = \eta = g(f), \tag{3.1}$$

where η is the linear predictor and $g(\cdot)$ is the link function.

The output parameter GRADIENTS is not supplied by RKEEP if a GLM has been fitted. INFKEEP therefore calculates GRADIENTS using the relation

$$\frac{\partial f}{\partial \beta_i} = X_i \frac{df}{d\eta}, \tag{3.2}$$

for i from 1 to p. INFKEEP extracts X_i from the DESIGN output parameter of RKEEP, and calculates $df/d\eta$ using the ITERATIVEWEIGHTS output parameter of RKEEP and the identity

$$W(f) = [V(f)]^{-1}\left(\frac{df}{d\eta}\right)^2, \tag{3.3}$$

where $W(f)$ is the iterative weight, expressed as a function of f, and $V(\cdot)$ is the variance function. (See Section 8.5.6 of the Genstat manual [6] or Section 2.5.1 of McCullagh and Nelder, 1983 [7].) INFKEEP calculates $V(f)$ using an expression chosen according to the DISTRIBUTION option of RKEEP, and deduces the sign of $df/d\eta$ from the LINK and EXPONENT options of RKEEP.

Considering the output parameter INVERSE, we note that, for a GLM, the ψ-function is the derivative, with respect to β, of $-2 \times$ log conditional likelihood of Y given X, and that

the i,j'th element of the ψ'-function is therefore

$$\begin{aligned}[\psi'(Y,X,\beta)]_{ij} =& 2\frac{\partial}{\partial\beta_j}\left\{(f-Y)[V(f)]^{-1}\frac{df}{d\eta}X_i\right\}\\ =& 2(f-Y)\frac{\partial}{\partial\beta_j}\left\{[V(f)]^{-1}\frac{df}{d\eta}X_i\right\}+2\left\{[V(f)]^{-1}\frac{df}{d\eta}X_i\right\}\frac{\partial}{\partial\beta_j}(f-Y)\\ =& 2(f-Y)\frac{\partial}{\partial\beta_j}\left\{[V(f)]^{-1}\frac{df}{d\eta}X_i\right\}+2W(f)X_iX_j. \qquad (3.4)\end{aligned}$$

(See Section 2.5.1 of McCullagh and Nelder (1983) [7].) If a GLM has been fitted, then the INVERSE parameter of RKEEP is the inverse sum of terms of the form

$$W(f)XX^T \qquad (3.5)$$

(in the present notation), and the terms (3.5) correspond to half the second term of the bottom line of (3.4).

McCullagh and Nelder note that, in the case of a canonical link function, the first term in the bottom line of (3.4) is zero for all Y, X and β, because the expression $[V(f)]^{-1}df/d\eta$ is constant in β. It follows that, if a GLM with a canonical link function has been fitted, the INVERSE delivered by RKEEP can be multiplied by $n/2$ to give an INVPSIPRIME that is equal to the inverse mean ψ'-function of (1.6), as is the case for nonlinear models fitted by the Newton-Raphson method. In the case of a non-canonical link function, if the model is "true" in the sense that the conditional mean value of Y given X is equal to $X^T\beta$ for the "true" value of β, then the first term of the bottom line of (3.4) has a mean value of zero for the "true" β, so the INVPSIPRIME calculated by INFKEEP is presumably still a consistent estimator for the true inverse mean ψ'-function of (1.3). Presumably, therefore, in the case of a nonlinear model, or a GLM with a canonical link, the influence curve and standard errors calculated by INFKEEP are asymptotically robust to heteroscedasticity, overdispersion, underdispersion, and also to some kinds of lack of fit, in which case the θ_F (or β_F) being estimated is the value of θ (or β) giving the best fit to the joint population distribution of the variables Y and X. In the case of a GLM with non-canonical link, robustness to lack of fit may be lost, but the present author has not had time to investigate how important this robustness is likely to be in practice.

4. Possible amendments

There are 2 ways, that have occurred to the present author, in which INFKEEP might possibly be improved:

1. Presently, when the covariance matrix of the multivariate influence function INFCUR is calculated, it is done by dividing the sum of squares and products matrix (with terms INFCUR[ ]) by the denominator $n-1$, where n is the sample number. Huber (1981) [4], in his section on studentisation, suggests that a better idea might be to use the denominator $n-p$, where p is the number of parameters.

2. Currently, if a GLM or nonlinear model has been fitted with a WEIGHTS option on the MODEL statement, INFKEEP defines the output parameters RHO and PSI to be the weighted ρ- and ψ-functions, derived by multiplying those implicit in (2.2) by WEIGHTS, and uses the weighted PSI to calculate INFCUR (by multiplying it by INVPSIPRIME). It might be a better idea to return RHO and PSI unweighted, while still using a weighted PSI for calculating INFCUR.

5. References

[1] Ruppert, D. *M*-estimators. In Kotz, S., and Johnson, N. L. (editors), *Encyclopedia of Statistical Sciences*, Wiley, New York, 1985. Vol. 5, pp. 443-449.

[2] Serfling, R. J. *Approximation theorems of mathematical statistics.* Wiley, New York, 1985.

[3] Silvey, S. D. *Statistical inference.* Chapman and Hall, London, 1975.

[4] Huber, P. J. *Robust statistics.* Wiley, New York, 1981.

[5] Reid, N. Influence functions. In Kotz, S., and Johnson, N. L. (editors), *Encyclopedia of Statistical Sciences*, Wiley, New York, 1985. Vol. 5, pp. 117-119.

[6] Genstat 5 Committee. *Genstat 5 Release 3 Reference Manual.* Clarendon Press, Oxford, 1993.

[7] McCullagh, P., and Nelder, J. A. *Generalized linear models.* Chapman and Hall, London, 1983.

Patron: Her Majesty The Queen

IACR-Rothamsted
Integrated Approach to Crop Research

IACR-Rothamsted
Harpenden, Hertfordshire AL5 2JQ
Telephone: (01582) 763133
International: +44 1582 763133
Fax: (01582) 760981/467116

Director:
Professor B J Miflin

Acting Head of Statistics Department:
Miss Janet Riley

7 July 1995

Dr R B Newson
Room 11
57 Warwick Road
Earl's Court
London SW5

Dear Roger

I have now, at last, been able to obtain a referee's report about your procedure INFKEEP, a copy of which I enclose.

The referee has pointed out several deficiencies in the description and notes that most of the problems for which it is intended can be solved in other ways. Consequently I do not think that I can justify including it in the Library itself. I am sorry not to be able to be more positive, but an alternative possibility for distributing the procedure might be to put it into the Statlib Genstat library that Peter Lane has just set up. However, you might want to revise at least the paper first – let me know.

I hope that all goes well for you.

With best wishes

Yours sincerely

Roger

Roger W Payne

peter.lane@bbsrc.ac.uk

Referee's Report on 'INFKEEP, a Genstat procedure for calculating influence functions and associated standard errors,' by Roger Newson.

The author's present address is required.

Why is it called INFKEEP apart from the fact that it uses the kept output structures from FIT? This does not seem to be a very evocative mnemonic.

The general style of the introduction is very different from that of the Genstat Procedure Library. While the text might comply with the requirements of a theoretical journal, for practical readers it should start from a different premise.

The paper starts with a formal description of Huber's M-estimators, which were introduced to justify certain procedures such as robust regression. The notation is no doubt consistent with that of Serfling, but not with that of the Genstat manual, so that the reader has to work hard to relate what is being said to more familiar concepts. For example the generalisation of the log likelihood is described as the mean of a function, when we would expect to be a sum of components, or just a function. Parameters and variables are described in the singular when they are nearly always vectors and matrices, and the influence function is called a curve, although it is multidimensional. The derivative of the function is given a different symbol, and the second derivative therefore appears as a first derivative, which seems to be unnecessarily complicated. Having described the method as most general it then states that the observations are usually assumed independent and identically distributed, although in the test examples generalized linear models are used which imply that the variances are not identical.

Few readers will have heard of M-estimators as they are at most a footnote in some textbooks. What is required is first of all a description of the reason why it may be of interest to modify the standard procedures for estimating parameters, their standard errors and their confidence intervals. The main example given is that of overdispersion in the Poisson distribution: this problem is already catered for at the simplest level by use of the heterogeneity factor (which increases the variances but does not alter the estimates), or at a more sophisticated level by using the negative binomial distribution (to be provided as a direct option in the next release of Genstat) which modifies the parameter estimates as well. The problem of heteroscedasticity is catered for by defining a weight variate, which may be prescribed in advance or calculated as a function of parameters. The problem of outliers and influential points is catered for by the Genstat regression diagnostics, which allow the user to decide what to do about points so identified; the M-estimator philosophy seems to demand that we keep the points regardless but modify the analysis without identifying them explicitly.

The fact that the problems that gave rise to INFKEEP can be largely solved by other means does not mean that it should not be considered as an interesting procedure. However the description must first set out when and why it is necessary to consider modifying the standard output based on likelihood theory. The tests do not seem to illustrate situations where the procedure is necessary, so why not take a standard example from a Robustness text, or even the author's own examples on insect counts? The essential feature seems to be use of the sample influence curve elements, as defined by Cook and Weisberg.

I would recommend a theoretical description in which one starts with the simple least squares model $E(y)=A\theta$, $D^2(y)=I\sigma^2$, $C=A'A$, $\theta\hat{}=C^{-1}A'y$, $D^2(\theta\hat{})=C^{-1}\sigma^2$, $y\hat{}=AC^{-1}A'y = Hy$, $D^2(y\hat{})=H\sigma^2$ and $D^2(y-y\hat{})=(I-H)\sigma^2$. Then the influence vectors are rows of $C^{-1}A'$ which are fixed by the design matrix **A**. From this model proceed to weighted least squares introducing the general dispersion matrix $V\sigma^2$. Then generalized linear models follow through transformations of **A** and quadratic approximations to **V** depending on the distribution and the fitted parameters. Models non-linear in the expectation follow by replacing **A** by the first partial derivatives of expectations at the fitted value, while non-normal error distributions are expressed through local approximations to the matrix **V**. Finally the general description via M-estimators may be given, with the distinction between the sample influence values and the empirical influence values.

There remain several unanswered questions, however.

1) Why is the estimation of the parameter separate from the estimation of variances, when a change of optimisation criterion may affect the estimates considerably?

2) There is a danger in the use of robust regression in that too exclusive an attitude to large residuals may result in local optima of the criterion: the parameters first thought of define which points will be treated as outliers, whereas from a different starting point a different set of outliers may be found.

3) Mention is made of confidence limits, but it is only the variances that are calculated, so presumably symmetric confidence limits will be computed using the Normal or t-statistics. In generalized linear models and non-linear models the parameters are transformable, and often exhibit very asymmetric confidence limits. The computations necessary to determine these involve constrained optimisation techniques.

There are further problems with the text as submitted, which should be noted in any revision:

Page 1:

The reference to Ruppert should be to the Kotz and Johnson encyclopedia, rather than to the author of the particular article. References should be in alphabetical order of first author.

The influence curve is strictly speaking the continuous version of (1.3) in which an independent variable x may be varied continuously to generate influence values, rather than the finite set of points represented by the data. The word 'curve' does not go easily in later phrases such as 'the dispersion matrix of the curve'.

Page 2:

Halfway down we read about confidence limits, yet it is only the dispersion matrix that is defined.

Page 3:

line 9; 'regression save structure' should have capitals to distinguish it from the ordinary text.

line 12; 'excellent optimisation software' is no doubt meant as a kind tribute but is out of place in a formal description.

line -2: 'general' should be 'generalized'.

Page 4:

line 6; 'sample number' should be 'sample size'

Binomial distribution uses n (already defined as sample number) where it should read n_i.

The square-root of f_i should be signed according to whether the observation is greater or less than its expectation. Otherwise the algebra will not work.

Reference to 'particular selling point' is out of place. Some users feel that it would be better to allow derivatives to be supplied if it is convenient so to do.

Page 5:

Gauss-Newton and Newton-Raphson have hyphens.

'apparently' is out of place. The statement is true, and is well known as the key difference between Gauss-Newton and Newton-Raphson procedures.

Page 7:

line -2; 'the present author has not had time' is of no relevance to the reader. Just say that it is not known.

Page 8:

'2 ways' should be 'two ways'.

Omit 'that have occurred to the present author'. There are no other authors, and it may be assumed that everything in a paper not attributed to others are the views of the author.