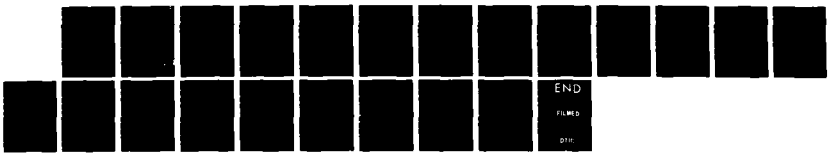
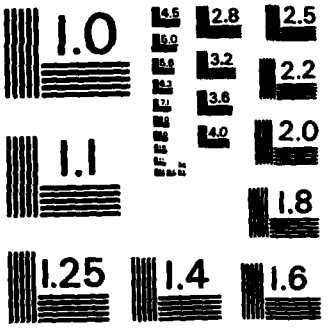
AD-A160 348 OPTIMALLY BOUNDED SCORE FUNCTIONS FOR GENERALIZED 1/1
LINEAR MODELS WITH APPL.. (U) NORTH CAROLINA UNIV AT
CHAPEL HILL DEPT OF STATISTICS L A STEFANSKI ET AL.
UNCLASSIFIED APR 85 AFOSR-TR-85-0866 F49620-82-C-0009 F/G 12/1 NL





MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS - 1963 - A

NONCLASSIFIED

SECURITY CLASSIFICATIO

AD-A160 348

TION PAGE

1a. REPORT SECURITY CL
Nonclassified

STRICTIVE MARKINGS

2a. SECURITY CLASSIFICATION AUTHORITY

3. DISTRIBUTION/AVAILABILITY OF REPORT
Approved for public release; distribution unlimited.

2b. DECLASSIFICATION/DOWNGRADING SCHEDULE

4. PERFORMING ORGANIZATION REPORT NUMBER(S)

5. MONITORING ORGANIZATION REPORT NUMBER(S)
AFOSR-TR- 85-0866

6a. NAME OF PERFORMING ORGANIZATION
Univ. of North Carolina

6b. OFFICE SYMBOL (If applicable)

7a. NAME OF MONITORING ORGANIZATION
Air Force Office of Scientific Research

6c. ADDRESS (City, State and ZIP Code)
Department of Statistics
Phillips Hall, 039A
Carolina Campus

7b. ADDRESS (City, State and ZIP Code)
Bolling AFB D.C. 20332

8a. NAME OF FUNDING/SPONSORING ORGANIZATION
AFOSR

8b. OFFICE SYMBOL (If applicable)

9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER
F49620 82 C 0009

8c. ADDRESS (City, State and ZIP Code)
Bolling Air Force Base
Washington, DC 20332

10. SOURCE OF FUNDING NOS.

PROGRAM ELEMENT NO.	PROJECT NO.	TASK NO.	WORK UNIT NO.
61102F	2304	A5	

11. TITLE (Include Security Classification)
"Optimally Bounded Score Functions for Generalized Linear Models with Applications to Logistic Regression"

12. PERSONAL AUTHOR(S)
Stefanski, Leonard A., Carroll, Raymond J. and Ruppert, David

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT (Yr., Mo., Day)	15. PAGE COUNT
technical	FROM 9/84 TO 8/85	April 1985	18

16. SUPPLEMENTARY NOTATION

17. COSATI CODES		
FIELD	GROUP	SUB. GR.

18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
Bounded influence, Generalized linear models, Influential points, Logistic regression, Outliers, Robustness.

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

We study optimally bounded score functions for estimating regression parameters in a generalized linear model. Our work extends results obtained by Krasker & Welsch (1982) for the linear model and provides a simple proof of Krasker and Welsch's first order condition for strong optimality. The application of these results to logistic regression is studied in some detail with an example given comparing the bounded influence estimator with maximum likelihood.

Publication keywords: outliers; robustness; influential points.

DTIC FILE COPY

DTIC ELECTE OCT 15 1985 E

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT
UNCLASSIFIED/UNLIMITED ☒ SAME AS RPT. ☐ DTIC USERS ☐

21. ABSTRACT SECURITY CLASSIFICAT
Nonclassified

22a. NAME OF RESPONSIBLE INDIVIDUAL	22b. TELEPHONE NUMBER (Include Area Code)	22c. OFFICE SYMBOL
Brian W. Woodruff, MAJ, USAF	(202) 767-5026	AFOSR/NM

DD FORM 1473, 83 APR EDITION OF 1 JAN 73 IS OBSOLETE.

NONCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE

MST

AFOSR-TR- 85-0866

OPTIMALLY BOUNDED SCORE FUNCTIONS FOR GENERALIZED LINEAR MODELS WITH APPLICATIONS TO LOGISTIC REGRESSION

by

Leonard A. Stefanski
Department of Economic and Social Statistics
Cornell University
Ithaca, New York 15853

Raymond J. Carroll
Department of Statistics
University of North Carolina
Chapel Hill, North Carolina 27514

David Ruppert
Department of Statistics
University of North Carolina
Chapel Hill, North Carolina 27514

85 10 11 131

SUMMARY

We study optimally bounded score functions for estimating regression parameters in a generalized linear model. Our work extends results obtained by Krasker & Welsch (1982) for the linear model and provides a simple proof of Krasker and Welsch's first order condition for strong optimality. The application of these results to logistic regression is studied in some detail with an example given comparing the bounded influence estimator with maximum likelihood.

Some key words: Bounded influence; Generalized linear models; Influential points; Logistic regression; Outliers; Robustness.

1. INTRODUCTION

In a generalized linear model (McCullagh & Nelder, 1983, Ch. 2) a response variable Y and covariate vector X are related via a conditional density of the form

$$f(y|x) = \exp\{(y-h(x^T\theta))q(x^T\theta)\omega(x)/\sigma + C(y,\sigma)\} \ .$$

The functions $h(\cdot)$ and $q(\cdot)$ are subject to certain restrictions, θ is a vector of regression parameters, σ is a scale parameter and $\omega(x)$ is a known weight function. In this paper we study the problem of robustly estimating θ when $\omega(\cdot) \equiv 1$ and σ is known. Models of this type include logistic and probit regression, Poisson regression, and certain models used in modeling lifetime data. In the case where $\omega(\cdot) \equiv 1$ but σ is unknown the methods presented in Section 2 are still applicable with some modification to allow for joint estimation of σ, c.f. Krasker & Welsch (1982).

Our motivation for seeking robust estimators is the same as that encountered in the context of linear model -- maximum likelihood estimation is very sensitive to outlying data. For the case of logistic regression, Pregibon (1981, 1982) has documented the nonrobustness of the maximum likelihood estimator and expounded the benefits of diagnostics as well as robust or resistant fitting procedures.

Much of the work on robust estimation concerns finding estimators which sacrifice little efficiency at the assumed model while providing protection against outliers and model violations. We follow this course finding bounded influence estimators minimizing certain functionals of the asymptotic covariance matrix. Related work includes that of Hampel (1978), Krasker (1980), and Krasker & Welsch (1982).

Two important issues when fitting models to data are (i) identification of outliers and influential cases and (ii) accommodation of these

observations. Frequently when influential cases are present, the fitted model is not representative of the bulk of the data. To rectify this, one can simply delete influential cases and refit via standard methods, but this approach lacks a theory for inference and testing; the effects of case deletion upon the distributions of estimators is not well understood, even asymptotically.

The robust techniques studied here provide a method of accommodating anomalous data. They allow continuous downweighting of influential cases and are amenable to asymptotic inference. Also, together with more direct diagnostics, residuals and weights from a bounded influence fit can be used to detect exceptional observations.

In Section 2 we present some general theory; this is specialized to the case of logistic regression in Section 3; proofs of theorems are given in an appendix.

2. THE GENERAL THEORY

2.1 The regression model

We study regression models in which the dependent variable Y and explanatory p-vector X have a density of the form

$$g(y,x;\theta_0) = f(y;x^T\theta_0)s(x). \tag{2.1}$$

The conditional density of Y given X=x is $f(y;x^T\theta_0)$ and depends on the unknown parameter θ_0 only through $x^T\theta_0$; s(x) is the marginal density of X. Expectation with respect to $g(y,x;\theta)$ is denoted by E_θ while $E_{\theta,x}$ indicates conditional expectation corresponding to $f(y;x^T\theta)$. Model (2.1) includes many generalized linear models (McCullagh & Nelder, 1983, Ch. 2).

Suppose (Y_i,X_i), (i=1,...,n) are independent copies of (Y,X). Under regularity conditions the maximum likelihood estimator of θ_0 satisfies

$$\sum_{i=1}^{n} \psi(Y_i,X_i,\hat{\theta}_{ML}) = 0,$$

where $\dot{\ell}(y,x,\theta) = \partial/\partial\theta\{\log(f(y;x^T\theta))\}$, and $n^{\frac{1}{2}}(\hat{\theta}_{ML} - \theta_o)$ converges in distribution to a p-dimensional normal random variate with mean zero and covariance matrix $V(\theta_o) = E_{\theta_o}^{-1}\{\dot{\ell}(Y,X,\theta_o)\dot{\ell}^T(Y,X,\theta_o)\}$.

2.2 M-estimators and their influence curves

We will consider estimators $\hat{\theta}_\psi$ satisfying

$$\sum_{i=1}^{n} \psi(Y_i,X_i,\hat{\theta}_\psi) = 0 ,$$

for suitably chosen functions ψ from $R \times R^P \times R^P$ to R^P. We require that ψ be unbiased, i.e.,

$$E_\theta\{\psi(Y,X,\theta)\} = 0 . \tag{2.2}$$

Under regularity conditions (Huber, 1967), $\hat{\theta}_\psi$ is consistent and asymptotically normal with influence curve

$$IC_\psi(y,x,\theta) = D_\psi^{-1}(\theta)\psi(y,x,\theta) \tag{2.3}$$

where $$D_\psi(\theta) = -\partial/\partial\beta[E_\theta\{\psi(Y,X,\beta)\}]_{\beta=\theta} . \tag{2.4}$$

Write $\psi(\theta)$ and $\dot{\ell}(\theta)$ for $\psi(Y,X,\theta)$ and $\dot{\ell}(Y,X,\theta)$ respectively. Assuming that integration and differentation can be interchanged in (2.2) and (2.4) it is easy to show that

$$D_\psi(\theta) = E_\theta\{\psi(\theta)\dot{\ell}^T(\theta)\}. \tag{2.5}$$

Now let

$$W_\psi(\theta) = E_\theta\{\psi(\theta)\psi^T(\theta)\} . \tag{2.6}$$

It then follows (Huber, 1967) that the asymptotic variance of $n^{\frac{1}{2}}(\hat{\theta}_\psi - \theta_o)$ is

$$V_\psi(\theta_o) = D_\psi^{-1}(\theta_o)W_\psi(\theta_o)(D_\psi^{-1}(\theta_o))^T.$$

For robustness we want IC_ψ to be bounded; for efficiency we want V_ψ to be small. In the next section we define a norm for IC_ψ and outline a

theory which suggests efficient bounded score functions.

2.3 A scalar measure of influence and an optimal score function

As a scalar measure of maximum influence we employ a definition of sensitivity introduced by Stahel in his Swiss Federal Institute of Technology Ph.D. thesis, and by Krasker & Welsch (1982). The *self standardized sensitivity* of the estimator $\hat{\theta}_\psi$ is defined as

$$s(\psi) = \sup_{(y,x)} \sup_{\lambda \neq 0} \frac{|\lambda^T IC_\psi|}{(\lambda^T V_\psi \lambda)^{\frac{1}{2}}} = \sup_{(y,x)} (IC_\psi^T V_\psi^{-1} IC_\psi)^{\frac{1}{2}}$$

$$= \sup_{(y,x)} (\psi^T W_\psi^{-1} \psi)^{\frac{1}{2}} . \qquad (2.7)$$

For a generalized linear model $s(\psi)$ has a natural interpretation in terms of the link function, e.g., in logistic regression $s(\psi)$ measures the maximum normalized influence of (y,x) on an estimated logit in that $\lambda^T IC_\psi$ is the influence curve for $\lambda^T \hat{\theta}_\psi$ and $\lambda^T V_\psi \lambda$ is the asymptotic variance of $\lambda^T \hat{\theta}$. Although this paper studies only the self standardized sensitivity we believe that useful estimators can also be obtained by bounding other measures of influence, such as fitted values.

For maximum likelihood $\psi = \ell$ and, in general, $s(\ell) = +\infty$. To obtain robustness we limit attention to only those estimators $\hat{\theta}_\psi$ for which

$$s(\psi) \leq b < \infty . \qquad (2.8)$$

Such an estimator is said to have bounded influence with bound b.

Consider the score function

$$\psi_{BI}(y,x,\theta) = (\ell - C)\min^{\frac{1}{2}}\{1, b^2/((\ell - C)^T B^{-1}(\ell - C))\} , \qquad (2.9)$$

where $\ell = \ell(y,x,\theta)$ and $C_{p\times 1} = C(\theta)$ and $B_{p\times p} = B(\theta)$ are functions of θ defined implicitly by the equations

$$E_\theta\{\psi_{BI}(y,x,\theta)\} = 0, \qquad B(\theta) = E_\theta\{\psi_{BI}\psi_{BI}^T\}. \tag{2.10}$$

With $C(\theta)$ and $B(\theta)$ so defined, ψ_{BI} is unbiased and $W_{\psi_{BI}}(\theta) = B(\theta)$, so that by (2.7), ψ_{BI} has bounded sensitivity.

The vector $C(\theta)$ and matrix $B(\theta)$ are analogous to robust multivariate location and scatter functionals for $\lambda(X,X,\theta)$, (Maronna, 1976). For sufficiently large b solutions $C(\theta)$ and $B(\theta)$ satisfying (2.10) exist, and as b tends to infinity these tend to zero and $E\{\lambda\lambda^T\}$ respectively. Equation (2.9) shows that ψ_{BI} is similar to a weighted maximum likelihood score with weights depending on the distance $(\lambda-C)^T B^{-1}(\lambda-C)$; as b tends to infinity the weighting factor tends to one and ψ_{BI} to λ.

For the normal theory linear model ψ_{BI} is the score function found by Krasker & Welsch (1982), who show that if there exists a score ψ_{opt} satisfying (2.2) and (2.8) which minimizes V_ψ in the strong sense of positive definiteness, i.e., $V_\psi - V_{\psi_{opt}} \geq 0$ for all ψ, then it must be of the form (2.9). That ψ_{BI} possesses similar optimality properties is seen in Corollary 1.1 below.

THEOREM 1. *If for a given choice of* $b > 0$ *equations* (2.10) *possess the solution* $(C(\theta), B(\theta))$, *then* ψ_{BI} *minimizes* $tr(V_\psi V_{BI}^{-1})$ *among all* ψ *satisfying* (2.2) *and*

$$\sup_{(y,x)} (IC_\psi^T V_{BI}^{-1} IC_\psi) \leq b^2 \quad . \tag{2.11}$$

With the exception of multiplication by a constant matrix, ψ_{BI} *is unique almost surely.*

Any score function ψ_{opt} for which $V_\psi - V_{\psi_{opt}} \geq 0$ for all ψ will be called strongly efficient; we now state the following corollary.

COROLLARY 1.1. *If there exists an unbiased, strongly efficient score* ψ_{opt} *satisfying* (2.8), *then* ψ_{opt} *is equivalent to* ψ_{BI} *whenever the latter is defined.*

Remarks 1. In Theorem 1 the conditions for optimality of ψ_{BI} depend on ψ_{BI} itself through V_{BI}^{-1}. This is somewhat disconcerting. Nevertheless ψ_{BI} does satisfy an optimality property and this result allows us to prove Corollary 1.1.

2. Working within the class of score functions of the form $A(y,x,\theta)\omega(y,x,\theta)$ where ω is a scalar weight function, Krasker & Welsch (1982) find the optimal form of ω. Theorem 1 and its corollary show that ψ_{BI} is optimal over a much larger class of functions and hence yield a technically stronger result than Krasker and Welsch's. Also our proof is somewhat simpler than Krasker and Welsch's.

3. Ruppert (1985) has shown that a strongly efficient score need not exist, in which case Corollary 1.1 is vacuous. In fact, we know of no case with $p \geq 2$ where a strongly efficient score has been shown to exist. However, the result given in Corollary 1.1 is still of interest; Ruppert (1985) uses it in his counter example.

4. The proofs of Theorem 1 and its corollary are presented in the appendix.

2.4 A one-step estimator

Write $\psi_{BI} = \psi_{BI}(y,x,\theta,B,C)$ to indicate dependence on B and C. Theorem 1 suggests the estimator $\hat{\theta}_{BI}$ obtained by solving

$$\sum_{i=1}^{n} \hat{\psi}_i(\hat{\theta}_{BI}) = 0$$

where $$\hat{\psi}_i(\theta) = \psi_{BI}(Y_i,X_i,\theta,\hat{B}(\theta),\hat{C}(\theta))$$

and $\hat{C}(\theta)$ and $\hat{B}(\theta)$ are defined implicitly by the equations

$$\sum_{i=1}^{n} E_{\theta,X_i}\{\hat{\psi}_i(\theta)\} = 0 \tag{2.12}$$

$$\hat{B}(\theta) = n^{-1}\sum_{i=1}^{n} E_{\theta,X_i}\{\hat{\psi}_i(\theta)\hat{\psi}_i^T(\theta)\} \ . \tag{2.13}$$

In the linear model (Krasker & Welsch, 1982) symmetry implies $\hat{C}(\theta) = 0$ so that finding $\hat{\theta}_{BI}$ is greatly simplified. For non-linear models solving for $\hat{\theta}_{BI}$ is much more difficult, so we suggest the following one-step procedure. Let $\tilde{\theta}$ be an initial root-n consistent estimator of θ_o. Compute $\hat{B}(\tilde{\theta})$ and $\hat{C}(\tilde{\theta})$ iteratively from (2.12) and (2.13). Define

$$\hat{\theta}_{BI}^{(1)} = \tilde{\theta} + n^{-1}\sum_{i=1}^{n} \hat{D}^{-1}(\tilde{\theta})\hat{\psi}_i(\tilde{\theta})$$

where $$\hat{D}(\theta) = n^{-1}\sum_{i=1}^{n} E_{\theta,X_i}\{\hat{\psi}_i(\theta)\lambda^T(Y_i,X_i,\theta)\} \ .$$

This construction is similar to Bickel's (1975) Type II one-step procedure. Under regularity conditions $\hat{\theta}_{BI}^{(1)}$ is consistent and asymptotically normal with covariance matrix $V_{BI}(\theta_o) = D_{BI}^{-1}(\theta_o)B(\theta_o)(D_{BI}^{-1}(\theta_o))^T$, which is consistently estimated by $\hat{V} = \hat{D}^{-1}(\tilde{\theta})\hat{B}(\tilde{\theta})(\hat{D}^{-1}(\tilde{\theta}))^T$. To preserve finite sample robustness we suggest that $\tilde{\theta}$ also be resistant to outliers.

3. APPLICATION TO LOGISTIC REGRESSION

3.1 The logistic model

Logistic regression is a special case of model (2.1) in which Y is an indicator variable such that

$$P(Y=1|X=x) = F(x^T\theta_o), \quad F(t) = 1/(1+\exp(-t)).$$

The general applicability of this form of binary regression is discussed by Berkson (1951), Cox (1970), and Efron (1975). The likelihood score is $\lambda(y,x,\theta) = (y-F(x^T\theta))x$ and the maximum likelihood estimator is consistent

and an asymptotically normal with covariance matrix $V(\theta_o) = E_{\theta_o}^{-1}\{F^{(1)}(X^T\theta_o)XX^T\}$ where $F^{(1)}(t) = (d/dt)F(t)$.

3.2 Constructing the one-step estimator for the logistic model.

The first step in computing $\hat{\theta}_{BI}^{(1)}$ entails finding an easily computed, robust, root-n consistent estimator $\tilde{\theta}$. We find an optimal score function from among the class,

$$M = \{\psi:\psi(y,x,\theta) = (y-F(x^T\theta))\omega(x,\theta)\}$$

where $\omega(\cdot,\cdot)$ is a p-vector valued function of x and θ but not y. The advantage, in terms of computational simplicity, of restricting attention to score functions in M is that condition (2.2) is automatically satisfied and it is not necessary to estimate a robust location functional.

The estimator we propose, and call a bounded leverage estimator, corresponds to the score

$$\psi_{BL} = (y - F(x^T\theta))x \min^{\frac{1}{2}}\{1,b^2/(m^2(x^T\theta)x^TQ^{-1}(\theta)x)\},$$

where $Q_{p\times p} = Q(\theta)$ is an implicitly defined function of θ satisfying

$$Q(\theta) = E_\theta\{F^{(1)}(X^T\theta)XX^T\min\{1,b^2/(m^2(X^T\theta)X^TQ^{-1}X)\}\}, \qquad (2.14)$$

and $m(\cdot)$ is the function $m(t) = \max(F(t), 1-F(t))$. In L. A. Stefanski's University of North Carolina Ph.D. thesis it is shown that in order for (2.14) to possess a solution $Q > 0$, it is necessary that

$$b^2 > p / E_\theta\{F^{(1)}(X^T\theta)/m^2(X^T\theta)\}. \qquad (2.15)$$

Condition (2.15) is generally not sufficient however. Note that with Q satisfying (2.14), $W_{\psi_{BL}} = Q$ and by (2.7), $\hat{\theta}_{BL}$ has bounded influence.

We are able to restrict attention to only those ψ in M and still obtain bounded influence simply because the absolute residual $|y - F(x^T\theta)|$ is bounded. However, ψ_{BL} takes a pessimistic view in downweighting observations in accordance with their maximum potential influence determined by their position in the design space and by θ. The term leverage is often used to denote potential influence (Cook & Weisburg, 1983) and

hence the name bounded leverage. Potential influence is often far greater than the actual influence when the observation is well fit by the model. Although downweighting such points results in a loss of efficiency for $\hat{\theta}_{BL}$ this will not affect the efficiency of our one-step estimator. Also, as the following results show, ψ_{BL} is the most efficient score in M.

THEOREM 2. *If for a given choice of* $b > 0$ *equation* (2.14) *possesses the solution* $Q > 0$ *then* ψ_{BL} *minimizes* $tr(V_\psi V_{BL}^{-1})$ *among all* ψ in M *satisfying*

$$\sup_{(y,x)} (IC_\psi^T V_{BL}^{-1} IC_\psi) \leq b^2 .$$

With the exception of multiplication by a constant matrix, ψ_{BL} *is unique almost surely.*

COROLLARY 2.1. *If there exists a strongly efficient score* ψ_{opt} *in* M, *then* ψ_{opt} *is equivalent to* ψ_{BL} *whenever the latter is defined.*

Remark 1. Proofs are similar to those of Theorem 1 and its corollary and will not be given.

2. The extent to which Theorem 2 generalizes to other regression models is limited, since it requires that $\lambda(y,x,\theta)$ be a bounded function of y.

Our initial estimator $\tilde{\theta}$ is obtained by solving

$$\sum_{i=1}^{n} (Y_i - F(X_i^T\tilde{\theta}))X_i \min^{\frac{1}{2}}\{1, b^2/(m^2(X_i^T\tilde{\theta})X_i^T\tilde{Q}^{-1}(\tilde{\theta})X_i)\} = 0$$

where $\tilde{Q}(\theta)$ satisfies

$$\tilde{Q}(\theta) = n^{-1} \sum_{i=1}^{n} F^{(1)}(X_i^T\theta)X_iX_i^T \min\{1, b^2/(m^2(X_i^T\theta)X_i^T\tilde{Q}^{-1}(\theta)X_i)\}.$$

For the one-step construction in Section 2.4 to work it is necessary that $\tilde{\theta}$ be root-n consistent. In Stefanski's Ph.D. thesis it is shown that $n^{\frac{1}{2}}(\tilde{\theta}-\theta_o)$ is asymptotically normal with covariance matrix $V_{BL}(\theta_o) = D_{BL}^{-1}(\theta_o)Q(\theta_o)(D_{BL}^{-1}(\theta_o))^T$ provided:

(i) b is sufficiently large,

(ii) $E\{||X||^2\} < \infty$,

(iii) $E\{F^{(1)}(X^T\theta)XX^T||X||^{-1}\}$ is positive definite,

(iv) $(\partial/\partial Q)E\{J(X,\theta,Q)\}$ is nonsingular where

$$J(X,\theta,Q) = Q - F^{(1)}(X^T\theta)XX^T\min\{1,b^2/(m^2(X^T\theta)X^TQ^{-1}X)\}.$$

The key assumptions are (iii) and (iv) which are similar to Assumption 7 of Krasker & Welsch (1982).

As an estimate of V_{BL} we use $\tilde{V} = \tilde{D}^{-1}(\tilde{\theta})\tilde{Q}(\tilde{\theta})(\tilde{D}^{-1}(\tilde{\theta}))^T$ where

$$\tilde{D}(\tilde{\theta}) = n^{-1}\sum_{i=1}^{n} E_{\tilde{\theta},X_i}\{\psi_{BL}(Y_i,X_i,\tilde{\theta},\tilde{Q}(\tilde{\theta}))\dot{L}^T(Y_i,X_i,\tilde{\theta})\}\ .$$

An algorithm for computing $\tilde{\theta}$ and $\hat{\theta}_{BI}^{(1)}$ for logistic regression models appears in Stefanski's Ph.D. thesis. To fully specify the algorithm one must determine the bound b. For $\tilde{\theta}$ this was chosen as a constant multiple of $b(\tilde{\theta})$ where

$$b^2(\theta) = p/[n^{-1}\sum_{i=1}^{n}\{F^{(1)}(X_i^T\theta)/m^2(X_i^T\theta)\}],$$

see (2.15). For the examples in the next section we took the bound to be $(1.5)b(\tilde{\theta})$; this same bound was then used for the one-step estimator $\hat{\theta}_{BI}^{(1)}$. The choice $(1.5)b(\theta)$ was suggested by experience; it is sufficiently small to provide protection from extreme observations yet large enough to avoid computational problems.

3.3 Example.

We apply our results to data relating participation in the U.S. Food Stamp Program to various socioeconomic indicators. The data, which are available from the first author, were selected at random from a cohort of over 2000 elderly citizens. The covariates are, (i) tenancy, indicating home ownership; (ii) supplemental income, indicating whether some form of supplemental security income is received; and (iii) monthly income. In our sample of 150 there were 24 cases of participation.

The researcher who supplied these data had been using probit regression with monthly income entering linearly in the model. A fit of the logistic model with covariates tenancy, supplemental income, and (monthly income)/10 produced Table 1(a).

Apart from the constant C, $\hat{\theta}_{BI}$ is a weighted maximum likelihood estimator with weights $\omega_i = \min^{\frac{1}{2}}\{1, b^2/((\lambda_i - C)^T B^{-1}(\lambda_i - C))\}$, where $\lambda_i = \lambda(Y_i, X_i, \theta)$, see equation (2.9). Estimated weights, $\hat{\omega}_i$, less than one indicate influential or ill-fitting observations. For the analysis in Table 1(a) $\hat{\omega}_{40} = 0.69$, $\hat{\omega}_{66} = 0.40$, $\hat{\omega}_{95} = 0.98$ and $\hat{\omega}_{109} = 0.62$ were the only weights less than one. Since these observations correspond to the four largest incomes among those receiving food stamps a transformation of income is indicated.

In Table 1(b) we present the analysis with log(monthly income + 1) replacing (monthly income)/10. This transformation substantially reduces the leverage of large income values but increases the leverage of small income values. For this model the bounded influence estimator downweighted only two observations with $\hat{\omega}_5 = 0.21$ and $\hat{\omega}_{66} = 0.76$. Case 66 has the largest income among those participating while case 5 has the smallest income among those not participating. Apparently cases #5 and #66 are influencing the maximum likelihood fit; this is indicated to a great extent by the bounded influence analysis and even more so by the maximum likeli-

hood fit with the two outlying cases removed.

An advantage of robust methods over maximum likelihood is that residual plots are more reliable for uncovering outliers. This is illustrated in Figure 1. Residuals (Cox (1970), p. 96; Pregibon (1981)) are plotted for both the maximum likelihood and bounded influence fits.

3.4 Conclusions.

Our bounded influence procedure provides a method of fitting meaningful models in the presence of anomalous data. Since similar models can be obtained by diagnosing and deleting outliers it is worth emphasizing that, unlike the method of case deletion, robust methods are amenable to asymptotic inference; this feature is important whenever hypothesis testing or confidence regions are objectives.

Robust procedures also supply useful diagnostic tools for model building. Variable selection, as well as estimation, can be influenced by anomalous data; Pregibon (1982) cites such an example. Often robust methods suggest variables appropriate for modeling the bulk of the data which would otherwise go undetected in a standard maximum likelihood analysis. Conversely, with non-resistant fitting, a variable might be used in the model simply to accommodate a single outlier. In addition to variable selection, the weights and residuals from a robust fit provide useful supplements to more direct diagnostics. For example, with the food stamp data, an analyst, seeing the impact of case five, might question the validity of that observation or the appropriateness of the model over the full range of incomes.

The research of the first two authors was supported by the Air Force Office of Scientific Research while that of the third author was supported

by the National Science Foundation. We also acknowledge the helpful comments of the referees.

APPENDIX

Proofs of Theorem 1 and Corollary 1

Theorem 1 is a generalization of Appendix A in Hampel (1978) and the proof given here uses techniques found in Kraaker (1980).

Proof of Theorem 1. Let ψ be any competitor to ψ_{BI}. Without loss of generality assume that $\psi = IC_{\psi}$, i.e. that ψ is in canonical form in the sense of Hampel (1974). This is equivalent to assuming

$$E_{\theta}\{\psi(Y,X,\theta)\lambda^{T}(Y,X,\theta)\} = I_{p\times p}, \tag{A.1}$$

and implies $V_{\psi}(\theta) = E_{\theta}\{\psi(Y,X,\theta)\psi^{T}(Y,X,\theta)\}$.

Now write λ for $\lambda(Y,X,\theta)$ and ψ for $\psi(Y,X,\theta)$. If ψ satisfies (A.1) and (2.2) then

$$E_{\theta}\{(D_{BI}^{-1}(\lambda-C) - \psi)(D_{BI}^{-1}(\lambda-C) - \psi)^{T}\} =$$

$$D_{BI}^{-1}E_{\theta}\{(\lambda-C)(\lambda-C)^{T}\}(D_{BI}^{-1})^{T} - D_{BI}^{-1} - (D_{BI}^{-1})^{T} + V_{\psi}(\theta).$$

Therefore $\mathrm{tr}(V_{\psi}V_{BI}^{-1})$ is, neglecting an additive constant independent of ψ, proportional to

$$E_{\theta}\{(D_{BI}^{-1}(\lambda-C) - \psi)^{T}V_{BI}^{-1}(D_{BI}^{-1}(\lambda-C) - \psi)\}. \tag{A.2}$$

Define $\phi = V_{BI}^{-\frac{1}{2}}\psi$; in terms of ϕ, (A.2) becomes

$$E_{\theta}\{\|\phi - V_{BI}^{-\frac{1}{2}}D_{BI}^{-1}(\lambda-C)\|^{2}\}. \tag{A.3}$$

Note that $\|\phi\|^{2} = \psi^{T}V_{BI}^{-1}\psi$ and thus subject to (2.11), equation (A.3) is minimized, as a function of ϕ, by

$$\phi = V_{BI}^{-\frac{1}{2}}D_{BI}^{-1}(\lambda-C)\min^{\frac{1}{2}}\{1, b^{2}/((\lambda-C)^{T}D_{BI}^{-1}V_{BI}^{-1}(D_{BI}^{-1})^{T}(\lambda-C))\}. \tag{A.4}$$

Condition (A.1) insures that ϕ is unique almost surely. Equations (2.3), (2.5), (2.6), and (2.10) imply $D_{BI}^{-1}V_{BI}^{-1}(D_{BI}^{-1})^{T} = B^{-1}$ thus in terms of ψ, (A.4) becomes $\psi = D_{BI}^{-1}\psi_{BI}$ proving the theorem. //

Proof of Corollary 1.1. Again assume that all scores are in canonical form and satisfy (2.2). Define

$$\mathcal{S} = \{\psi: \sup_{(y,x)} \psi^T V_\psi^{-1} \psi \le b^2\}, \quad \mathcal{S}_{BI} = \{\psi: \sup_{(y,x)} \psi^T V_{BI}^{-1} \psi \le b^2\} .$$

We must show that if there exists ψ_{opt} in $\mathcal{S}$ such that $V_{\psi_{opt}} \le V_\psi$ for all ψ in $\mathcal{S}$, then ψ_{opt} is equivalent to $D_{BI}^{-1}\psi_{BI}$. Clearly $D_{BI}^{-1}\psi_{BI}$ is in $\mathcal{S}$, thus by assumption $V_{\psi_{opt}} \le V_{BI}$. From this it follows that

$$\psi_{opt}^T V_{BI}^{-1} \psi_{opt} \le \psi_{opt}^T V_{\psi_{opt}}^{-1} \psi_{opt} \le b^2 ,$$

and hence ψ_{opt} is in $\mathcal{S}_{BI}$. Let $\mathcal{I} = \mathcal{S} \cap \mathcal{S}_{BI}$. The set $\mathcal{I}$ is nonempty; it contains $D_{BI}^{-1}\psi_{BI}$ and ψ_{opt}. For any ψ in $\mathcal{I}$ we know $V_{\psi_{opt}} \le V_\psi$ and hence

$$\mathrm{tr}(V_{\psi_{opt}} V_{BI}^{-1}) \le \mathrm{tr}(V_\psi V_{BI}^{-1})$$

for all ψ in $\mathcal{I}$. But Theorem 1 proves that $D_{BI}^{-1}\psi_{BI}$, when defined, is the almost everywhere unique minimizer of $\mathrm{tr}(V_\psi V_{BI}^{-1})$ among all ψ in $\mathcal{I}$. The equivalence of ψ_{opt} and ψ_{BI} follows. //

References

Berkson, J. (1951). Why I prefer logists to probits. *Biometrics* 7, 327-39.

Bickel, P. A. (1975). One-step Huber estimates in the linear model. *J. Am. Statist. Assoc.* 70, 428-34.

Cook, D. R. & Weisberg, S. (1983). Comment on "Minimax Aspects of Bonded-Influence Regression" by Huber. *J. Am. Statist. Assoc.* 78, 74-5.

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

Efron, B. (1975). The efficiency of logistic regression compared to normal discriminant analysis. *J. Am. Statist. Assoc.* 70, 892-8.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *J. Am. Statist. Assoc.* 69, 383-94.

Hampel, F. R. (1978). Optimally bounding the Gross-Error-Sensitivity and the influence of Position in factor space. *1978 Proceedings of the ASA Statistical Computing Section*, 59-64.

Huber, P. J. (1967). The behavior of maximum likelihood estimates under non-standard conditions. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statstics and Probability*. 1, Ed. L. M. LeCam and J. Neyman, pp. 221-33, University of California Press.

Krasker, W. S. (1980). Estimation in linear regression models with disparate data points. *Econometrics* 48, 1333-46.

Krasker, W. S. & Welsch, R. E. (1982). Efficient bounded influence regression estimation using alternative definitions of sensitivity. *J. Am. Statist. Assoc.* 77, 595-605.

Maronna, R. A. (1976). Robust M-estimators of multivariate location and scatter. *Ann. Statist.* 4, 51-67.

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

Pregibon, D. (1981). Logistic regression diagnostics. *Ann. Statist.* 9, 705-24.

Pregibon, D. (1982). Resistant fits for some commonly used logistic models with medical applications. *Biometrics* 38, 485-98.

Ruppert, D. (1985). On the bounded-influence regression estimator of Krasker and Welsch. *J. Am. Statist. Assoc.* 80, 205-08.

Table 1 (a). Estimates for the logistic regression model with covariates tenancy, supplemental income, and (monthly income)/10; p-values in parentheses.

	intercept	tenancy	supplemental income	(monthly income)/10
$\hat{\theta}_{ML}$	-0.34 (0.5287)	-1.76 (0.0009)	0.78 (0.1259)	-0.01 (0.1122)
$\tilde{\theta}$	-0.16 (0.7872)	-1.75 (0.0014)	0.77 (0.1360)	-0.02 (0.0826)
$\hat{\theta}^{(1)}_{BI}$	-0.20 (0.6006)	-1.76 (0.0012)	0.78 (0.1300)	-0.02 (0.0922)

Table 1 (b). Estimates for the logistic regression model with covariates tenancy, supplemental income, and log(monthly income + 1); p-values in parentheses.

	intercept	tenancy	supplemental income	log(monthly income + 1)
$\hat{\theta}_{ML}$	0.93 (0.5681)	-1.85 (0.0005)	0.90 (0.0737)	-0.33 (0.2228)
$\tilde{\theta}$	4.14 (0.1030)	-1.81 (0.0007)	0.75 (0.1444)	-0.86 (0.0430)
$\hat{\theta}^{(1)}_{BI}$	4.02 (0.1100)	-1.81 (0.0006)	0.76 (0.1416)	-0.84 (0.0465)
$\hat{\theta}^{*}_{ML}$	6.88 (0.0160)	-2.02 (0.0004)	0.76 (0.1586)	-1.33 (0.0062)

* With cases #5 and #66 excluded.

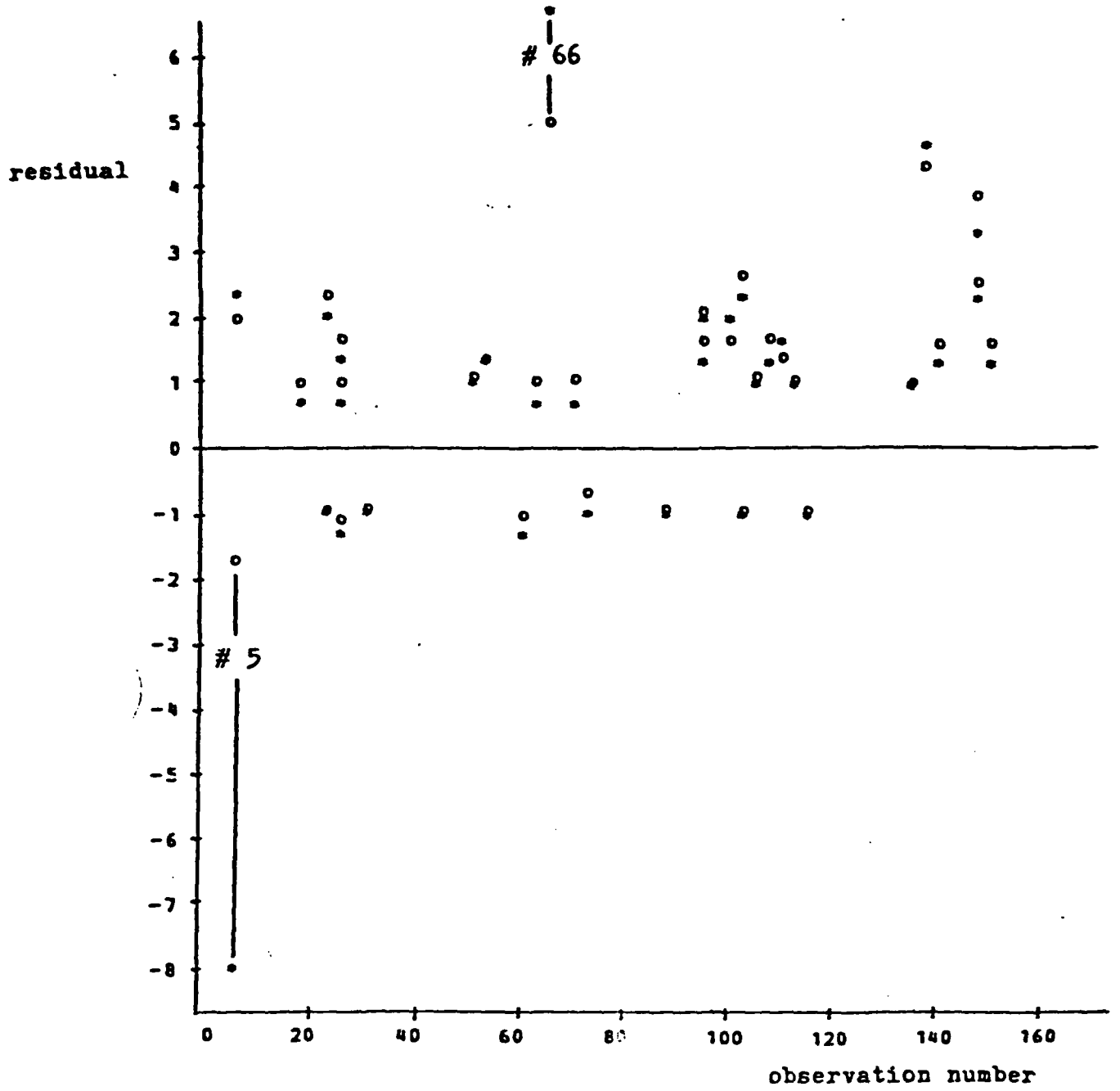


Figure 1. Residual Plots for FS Data. Maximum likelihood residuals are indicated by circles 'o'; residuals from the bounded influence fit by asterisks '*'. For both estimation procedures residuals are defined as in Cox (1970), p. 96. Negligible residuals have been omitted for clarity.

END

FILMED

11-85

DTIC