BOUNDING INFLUENCE AND LEVERAGE

IN LOGISTIC REGRESSION

by

Leonard A. Stefanski
Department of Economic and Social Statistics
Cornell University
Ithaca, New York 14853

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

Summary

Logistic regression is very sensitive to discordant data, see Pregibon (1981). Resistant fitting procedures, though clearly desirable, have not received their due attention. This paper studies two bounded influence estimators similar in spirit to the Krasker-Welsch (1982) estimator for linear regression. In motivating the bounded influence procedures some theoretical optimality results for Krasker-Welsch type estimators are obtained. These results are not specific to logistic regression and generalize to other statistical models, in particular the class of generalized linear models. The proposed estimation procedures are illustrated with examples.

Some Key Words: Binary regression; Bounded influence; Generalized linear models; Influential points; Logistic Regression; Outliers; Robustness.

1. Introduction

This article concerns robust estimation in the logistic regression model

$$\Pr(y_i = 1 \mid x_i) = F(x_i^T \theta) \quad i = 1, \ldots, N,$$
$$F(t) = 1/(1 + \exp(-t)). \tag{1.1}$$

The response variable y_i indicates the occurrence or nonoccurrence of some event, x_i is a p-vector of predictor variables, and θ is a p-vector of unknown parameters. The general applicability of model (1.1) is discussed by Berkson (1951), Cox (1970), and Efron (1975). The usual method of estimation is maximum likelihood and entails solving the equations

$$\sum_1^N (y_i - F(x_i^T \hat{\theta}))x_i = 0$$

for $\hat{\theta}_{ML}$. Under regularity conditions,

$$N^{\frac{1}{2}} (\hat{\theta}_{ML} - \theta) \xrightarrow{L} N(0, V),$$

where (1.2)

$$V^{-1} = \lim_N N^{-1} \sum_1^N F^{(1)}(x_i^T \theta)\, x_i\, x_i^T .$$

In (1.2) $F^{(1)}(\cdot)$ denotes the first derivative of $F(\cdot)$. As in the special case of Bernoulli trials, the asymptotic approximations are

generally reliable only when NP^* is large where $P^* = \min(P, 1 - P)$ and P is the sample proportion of positive responses. For this reason NP^* is sometimes referred to as the effective sample size.

Much like the least squares estimate in linear regression, $\hat{\theta}_{ML}$ is sensitive to aberrant observations. Pregibon (1981, 1982) has documented the nonrobustness of the maximum likelihood estimate and expounded the benefits of diagnostics as well as robust or resistant fitting procedures. This paper introduces and examines the performance of two bounded influence estimators similar in spirit to the linear regression estimator studied by Krasker and Welsch (1982).

In Section 2 some theoretical motivation for the estimators is presented. This material is not specific to logistic regression and generalizes to other statistical models, e.g. the class of generalized linear models, McCullagh and Nelder (1983). In addition to providing further justification for Krasker-Welsch type estimators our results admit a simple proof of the so-called Krasker-Welsch optimality result, see Theorem 1, Krasker and Welsch (1982). In Section 3 examples are used to illustrate the proposed methods. Section 4 concludes with proofs of the optimality results from Section 2.

Ideally a robust fit is responsive to the bulk of the observations and is not significantly influenced by any small subset of the data. In practice this can mean sacrificing the quality of the fit at a few points in order to obtain a better fit to the remaining data. This objective is well founded, especially if there is reason to suspect gross errors in the data and/or inadequacy of the model over some regions of the design space. With binary data, gross errors in

response variables can arise as a result of faulty data transcription or the inability to consistently assess the true response, e.g. a medical diagnosis is often subject to two types of error. Problems with gross errors in the predictors and model deficiencies are analogous to those encountered in linear regression. In the event that model (1.1) is severely inadequate one might choose to employ an expanded version such as proposed by Aranda-Ordaz (1981) or Guerrero and Johnson (1982). Also, if the predictor variables are subject to measurement error, rather than occasional gross errors, then errors-in-variables methods should be used (Carroll, Spiegelman, Lan, Bailey, and Abbott, 1984; Stefanski, 1983). However, even expanded models are sensitive to extreme observations and the simpler model (1.1) is the appropriate starting point for a study of robust methods.

There are two issues central to the problem of fitting models in the presence of anomalous data. These are (i) identification of poorly fit and/or influential cases and (ii) accommodation of these observations in the final analysis. For standard fitting procedures, e.g., least squares or maximum likelihood, identification of special cases can usually be accomplished with appropriate diagnostic techniques; for logistic regression see Pregibon (1981).

Frequently when influential cases are present, the fitted model is not representative of the bulk of the data. To rectify this situation one might simply choose to delete the influential cases from the sample and refit the model, via standard methods, to the remaining data thus obtaining a type of robust fit. Employed in this fashion

diagnostic methods also provide an ad hoc method of accommodation. The principle shortcoming of this approach is the lack of theory for inference and testing. The effects of data deletion upon the distributions of estimators and test statistics is not well understood, even asymptotically.

The robust techniques studied in this paper provide a method of accommodating anomalous data in the fitting process. They allow for continuous downweighting of influential cases and are amenable to asymptotic inference. Also the bounded influence weights and residuals can be employed as global diagnostics for detecting exceptional observations.

2. Bounding Influence and Leverage.

Although the discussion in this section focuses on logistic regression, the results presented extend to other settings. In particular Theorem 1 holds for rather general parametric families while the discussion culminating in Theorem 2 applies to general binary regression models.

Consider the class of estimates $\hat{\theta}_\psi$ defined by equations of the form

$$\sum_1^N \psi(y_i, x_i, \hat{\theta}_\psi) = 0 .$$

ψ is allowed to depend on N although the additional notation is usually suppressed. If $\psi(y, x, \theta) = \ell(y, x, \theta) \triangleq (y - F(x^T\theta))x$, then $\hat{\theta}_\psi$ is the maximum likelihood estimate. ψ will be called an unbiased score if $E_\theta(\psi(y_i\ x_i, \theta) = 0$ or more generally if

$$\sum_1^N E_\theta(\psi(y_i, x_i, \theta)) = 0 \quad . \tag{2.1}$$

Define, for any ψ,

$$D_{\psi}(\theta) \triangleq N^{-1} \sum_{1}^{N} E_{\theta}(\psi(y_i, x_i, \theta) \lambda^T (y_i, x_i, \theta)), \quad (2.2)$$

$$W_{\psi}(\theta) \triangleq N^{-1} \sum_{1}^{N} E_{\theta} (\psi(y_i, x_i, \theta) \psi^T (y_i, x_i, \theta)) . \quad (2.3)$$

When ψ is unbiased one expects, under appropriate regularity conditions, that as $N \to \infty$

$$N^{\frac{1}{2}} (\hat{\theta}_{\psi} - \theta - N^{-1} \sum_{1}^{N} D_{\psi}^{-1} (\theta) \psi(y_i, x_i, \theta)) \xrightarrow{P_{\theta}} 0, \text{ and} \quad (2.4)$$

$$N^{\frac{1}{2}} (\hat{\theta}_{\psi} - \theta) \dot{\sim} N(0, V_{\psi}(\theta)), \text{ where} \quad (2.5)$$

$$V_{\psi}(\theta) = D^{-1}_{\psi} (\theta) W_{\psi}(\theta) D^{-T}_{\psi}(\theta) . \quad (2.6)$$

Although our data are intrinsically independent but nonidentically distributed, equations (2.5) and (2.6) justify calling

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

the influence curve of the estimate $\hat{\theta}_{\psi}$. This is analogous to Hampel's (1968, 1974) definition of influence in independently identically distributed models. IC_{ψ} might also be called a conditional influence curve since all calculations are performed conditionally on the observed predictors.

As a scalar measure of maximum influence we employ a definition of sensitivity introduced by Stahel (1981) and Krasker and Welsch (1982). The self standardized sensitivity of the estimator $\hat{\theta}_{\psi}$ is defined as

$$s(\psi) = \sup_{(y,x)} (IC^T_{\psi} (y, x, \theta) V^{-1}_{\psi}(\theta) IC_{\psi}(y, x, \theta))^{\frac{1}{2}} .$$

In the definition of $s(\psi)$ for logistic regression the supremum is taken over the set $\{(y, x): y = 0, 1; x \in R^p\}$. It measures the maximum normalized influence of an observation (y, x) on an estimated logit in the sense that

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

$\lambda^T IC_\psi$ is the influence curve for $\lambda^T \hat{\theta}_\psi$, and $\lambda^T V_\psi(\theta)\lambda$ is the asymptotic variance of $\lambda^T \hat{\theta}_\psi$. Employing definitions (2.2), (2.3) and (2.6) we get an equivalent and useful expression for $s(\psi)$ in the form of

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

For maximum likelihood $\psi = \ell$ and $s(\ell) = +\infty$. 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 will be said to have bounded influence with bound b. Although this paper studies the self standardized sensitivity we believe that useful estimators can also be obtained by bounding other measures of influence, such as fitted values. Work on this is in progress.

There are generally an infinite number of estimators satisfying (2.1) and (2.8). Naturally we are interested in choosing ψ efficiently, i.e. so that V_ψ is small in some reasonably defined sense. Write ℓ for the likelihood $\ell(y, x, \theta)$. Our first bounded influence estimator corresponds to the score

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

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

$$\sum_{1}^{N} E_\theta \left(\psi_{BI}\left(y_i, x_i, \theta\right)\right) = 0 \,,$$

$$B = N^{-1} \sum_{1}^{N} E_\theta\left(\psi_{BI}\left(y_i, x_i, \theta\right) \psi_{BI}^{T}\left(y_i, x_i, \theta\right)\right) \,. \tag{2.10}$$

This procedure generalizes the approach taken by Krasker and Welsch (1982) for the linear model. With $C(\theta)$ and $B(\theta)$ so defined ψ_{BI} is unbiased and by equation (2.3) $W_{BI} \triangleq W_{\psi_{BI}} = B$. This together with (2.7) implies (2.8), so that $\hat{\theta}_{BI}$ has bounded influence. Equations (2.10) will not possess solutions for all $b > 0$. Results of Maronna (1976) suggest that asymptotically $b^2 \geq p$, is a necessary and sufficient condition for the existence of $C(\theta)$ and $B(\theta)$ satisfying (2.10). In finite samples this is a rough guide at best.

In the case of linear regression Krasker and Welsch (1982) show that if there exists a score ψ satisfying (2.1) and (2.8) which minimizes the asymptotic covariance V_ψ in the strong sense of positive definiteness, then it must be of the form (2.9). The fact that ψ_{BI} possesses similar optimality properties for generalized linear models and in particular for logistic regression is obtained as a corollary to the following result.

Theorem 1. If for a given choice of $b > 0$, equations (2.10) possess the solution $(C(\theta), B(\theta))$ then ψ_{BI} minimizes trace $(V_\psi V_{BI}^{-1})$ among all ψ satisfying (2.1) and

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

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

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 and necessarily weakens the statement of the theorem. Nevertheless ψ_{BI} does satisfy an optimality property and has the additional property of invariance.

2. Subject to regularity conditions, versions of Theorem 1 and its corollary hold in general parametric settings such as generalized linear models. Also the proofs we give are a good deal simpler than Krasker and Welsch's original proof for the linear model.

3. Ruppert (1983) 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 (1983) used that result to show that no strongly optimal estimator exists in his example. Thus Theorem 1 represents the best known general optimality result for Krasker-Welsch type estimators.

4. The proofs of Theorem 1 and its corollary are similar to those for Theorem 2 and Corollary 2.2 below, which are presented in Section 4.

Although ψ_{BI} satisfies an optimality criterion, computational difficulties make it less than fully acceptable. A less complex estimate can be obtained for logistic regression by exploiting the form of the likelihood and restricting attention to score functions in the class,

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

where $\omega(\cdot, \cdot)$ is a p-vector valued function of x and θ but not y. Our second estimator, the bounded leverage estimator, corresponds to the choice

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

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

$$Q = N^{-1} \sum_{1}^{N} (F^{(1)}(x_i^T\ \theta)\ x_i\ x_i^T\ r\ (b^2/(m^2(x_i^T\ \theta)\ x_i^T\ Q^{-1}x_i))), \quad (2.11)$$

and $m(\cdot)$ is the function $m(t) = \max(F(t), 1 - F(t))$. In order for (2.11) to possess a solution $Q > 0$, it is necessary that

$$b^2 \geq p\ /(N^{-1} \sum_{1}^{N} (F^{(1)}(x_i^T\ \theta)/m^2(x_i^T\ \theta))\ . \quad (2.12)$$

Condition (2.12) is generally not sufficient however, see Stefanski (1983) for details.

Note that with Q satisfying (2.11),equation (2.3) implies $W_{BL} \triangleq W_{\psi_{BL}} = Q$ and (2.8) holds, so that $\hat{\theta}_{BL}$ has bounded influence.

We are able to restrict attention to only those ψ in $\mathcal{M}$ and still obtain bounded influence simply because the absolute residual

$|y - F(x^T \theta)|$ is bounded for all (y, x). However, as a consequence of working within $\mathcal{M}$, ψ_{BL} takes a pessimistic view of the data since it downweights observations in accordance with their maximum potential influence determined as a function of their position in the design space and of θ. The term leverage is often used to denote the potential for a case to be influential (Cook and Weisburg, 1983) hence the term bounded leverage. The potential influence is often greater than the actual influence especially when the observation is well fit by the model. Although downweighting such points results in a loss of efficiency, the loss is often slight since potentially influential cases contribute little to the logistic information matrix. Heuristically this is because the influence contains the factor $|y - F|$ and the information contains the factor $F^{(1)} = F(1 - F)$. Whenever the residual is large, F is necessarily near zero or one and hence $F^{(1)}$ is small. The examples of Section 3 provide evidence supporting this argument.

Just as ψ_{BI} is optimal among the unbiased scores, ψ_{BL} is optimal within $\mathcal{M}$.

Theorem 2. If for a given choice of $b > 0$ equation (2.17) possesses the solution $Q > 0$ then ψ_{BL} minimizes trace $(V_\psi V_{BL}^{-1})$ among all ψ in $\mathcal{M}$ satisfying

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

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

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

Remark 1. Proofs are given in Section 4.

The extent to which Theorem 2 generalizes to other regression models is limited, since it requires that the likelihood score, $\ell(y, x, \theta)$, be a bounded function of the response variable y. This requirement is quite restrictive but of course holds for all binary response models.

The optimality results of this section are meaningful only when the asymptotic behavior indicated in (2.4) and (2.5) prevails. Design conditions insuring consistency and asymptotic normality of the optimal estimators can be found in Stefanski (1983).

3. Examples

In this section examples are used to study the performance of the proposed estimators. Of particular interest are (i) their ability to provide a meaningful model in the presence of influential and/or outlying data, i.e. a model which is representative of the bulk of the data; and (ii) the extent to which the robust procedures can be used to construct portmanteau diagnostics for detecting anomalous data.

Difficulties computing $\hat{\theta}_{BI}$ prompted us to examine a one-step version starting from $\hat{\theta}_{BL}$ and a biased version in which the bias correction vector, $C(\theta)$, is set equal to zero in equations (2.9) and (2.10). In the examples the one-step differed little from $\hat{\theta}_{BL}$ so we opted to present only the results for the biased estimate, designated $\hat{\theta}_{BI}(C = 0)$, and $\hat{\theta}_{BL}$. A modified Newton-Raphson iteration was employed to compute the robust estimates.

For comparison, the maximum likelihood estimate and Pregibon's (1982) resistant estimate, $\hat{\theta}_{RST1}$, are included in the study. Where employed, asymptotic variances for $\hat{\theta}_{ML}$, $\hat{\theta}_{BI}$ $(C = 0)$ and $\hat{\theta}_{BL}$ are those suggested by equations (2.5) and (2.6) with appropriate modifications in the case of $\hat{\theta}_{BI}(C = 0)$. The asymptotic variance for $\hat{\theta}_{RST1}$ was estimated in the manner employed by Pregibon (1982). It should be noted that the resulting variance estimator is generally not consistent.

Measures of observed efficiency with respect to maximum likelihood, defined as

$$\text{Eff}(\psi) = \inf_{\lambda} \left(\frac{\lambda^T V_{ML}(\theta)\lambda}{\lambda^T V_{\psi}(\theta)\lambda} \right) \Bigg|_{\theta = \hat{\theta}_{\psi}},$$

are also given. The method of calculating variance $(\hat{\theta}_{RST1})$ is necessarily optimistic and this manifests itself in observed efficiencies which our work suggests are inflated by approximately 10-15 percent.

For $\hat{\theta}_{BI}(C = 0)$ and $\hat{\theta}_{BL}$ the bound b was chosen as multiples of $p^{\frac{1}{2}}$ and

$$LB(\theta) \triangleq \left(p/(N^{-1} \sum_{1}^{N} F^{(1)}(x_i^T \theta) / m^2 (x_i^T \theta)) \right)^{\frac{1}{2}}$$

respectively, see Section 2. In each case the multiplier was adjusted to achieve roughly 90% observed efficiency as defined above. The choice of 90% represents a tradeoff. Generally this was low enough to yield reasonably resistant estimates yet sufficiently high to avoid computational difficulties. In practice it is sometimes useful to employ more than one bound; a tight bound for influence diagnostics, a

loose bound for inference. Here too the choice of 90% represents a compromise in the interest of saving space. For $\hat{\theta}_{RST1}$ the bound was fixed at 1.345, see Pregibon(1982) .

For each data set, table entries include the parameter estimate, asymptotic z-statistics for testing significance of individual components, and the measure of efficiency described above.

Example 1: Skin Vaso-constriction (SVC) Data

These data have appeared elsewhere in the context of robustness, Pregibon (1981, 1982). The response indicates the occurrence of vaso-constriction in the skin of the digits and is regressed on the logarithms of the rate and volume of air inspired. In our work we took $Rate_{32} = .30$, see Pregibon (1981). Pregibon has shown that observations 4 and 18 are influential.

Table 1 contains the results of the analyses. The similarity of the three robust fits is somewhat surprising in that the various procedures downweight influential cases according to markedly different criteria. The resistant estimate downweighted observations 4, 18, and 24, assigning them weights of 0.44, 0.47, and 0.97 respectively. The bounded influence estimate, downweighted only observations 4 and 18 with corresponding weights of 0.38 and 0.44. Thus both these procedures systematically revealed the aberrant observations 4 and 18. From the discussion in Section 2 it should be evident that the bounded leverage weights provide limited diagnostic information. In this

example approximately 2/3 of the data were downweighted. More significant is the fact that cases 4 and 18 received weights of 0.33 and 0.38,which are similar to the Krasker-Welsch weights for the same observations.

Example 2: Coronary Heart Disease (CHD) Data

These data are from a study relating coronary heart disease to systolic blood pressure and cholesterol level. The predictors employed are log(systolic blood pressure - 75) and log(cholesterol). Of the 108 cases, twelve were diagnosed as having heart disease. A plot of the data appears in Figure 1.

Table 2 displays the results of our analyses. We employed Pregibon's (1981) diagnostics to assess the maximum likelihood fit. They suggested that observations 4 and 35 are influential as well as being poorly accounted for by the model. The bounded influence estimate downweighted observations 4, 35, and 71 with weights of 0.79, 0.46, and 0.82 respectively. The resistant estimate assigned weights of 1.00, 0.43, and 0.53 to the same three observations but also downweighted five additional points with weights ranging from 0.60 to 0.95. The points downweighted by the resistant procedure comprise seven of the twelve reported cases of heart disease. For this example the bounded leverge weights are somewhat informative. To give the reader a feel for weighting criterion, Figure 1 is overlayed with a contour plot of the weights corresponding to all possible predictors. This plot indicates those regions of the design space corresponding to potentially influential cases. For reference, the line corresponding to a 50% chance of response, as estimated by $\hat{\theta}_{BL}$, is superimposed on

the plot.

The most notable difference between the maximum likelihood and robust estimates occurs in the coefficient of log (cholesterol). The results suggest that observations 4 and 35 obscure the importance of log(cholesterol) as a predictor when maximum likelihood estimation is employed. The entries in column 4 of Table 2 confirm this suspicion. For this analysis the maximum likelihood estimate was computed after removing observations 4 and 35 from the sample. The increased significance of log(cholesterol) as a predictor is evident.

Example 3: Food Stamp (FS) Data

Our last example comes from a study relating whether or not one participates in the Federal Food Stamp Program to three indicators of socioeconomic status. Two of the covariates are dichotomous; tenancy indicates home ownership; supplemental income indicates whether or not an individual receives supplemental security income. The third covariate is monthly income. Income data are frequently heavy tailed and often reexpressed with an appropriate transformation, e.g. logarithmic. The 150 observations (24 participating, 126 nonparticipating) in our sample include one zero monthly income, case number 5. The transformation, log (monthly income + 1) neatly handles the heavy tails of the income variable but creates, somewhat artificially, an extreme design point corresponding to the zero income value. It transpires that, for this transformation, observation number 5 is also very influential. While other transformations are possible the model with covariates tenancy, supplemental income, and log(monthly income + 1)

provides a good setting for illustrating the bounded influence procedures.

Table 3 displays the results of our analyses. Residuals, as defined by Cox (1970), p. 96, and Pregibon (1981), are plotted in Figure 2 for both the maximum likelihood and bounded leverage fits. Negligible residuals have been omitted for clarity. The most interesting feature of the plot is case number 5. For maximum likelihood estimation this observation has sufficient leverage to bend the model in its direction and, as a consequence, does not have an extreme residual. The robust residuals, on the other hand, clearly indicate the uniqueness of case 5. Although the maximum likelihood residual plot is fooled by the high leverage point it should be noted that case 5, and to a lesser extent case 66, are indicated as influential by certain of Pregibon's (1981) diagnostic plots.

The bounded-influence weights of 0.21 and 0.76 for observations 5 and 66 respectively, and the results in Table 3 suggest that case number 5 severely affects the perceived significance of log(monthly income + 1) as a predictor. In this respect the two bounded influence models are more consistent with the goal of modeling the bulk of the data. However, even these models are sensitive to observation 5. Lowering the bound on the influence improved the situation somewhat, but numerical difficulties were encountered before the effect of case 5 could be substantially reduced. As this example illustrates the bounded influence procedures are not panacean; in certain situations it may still be necessary to reject extreme outliers.

For these data the resistant estimate failed to produce a meaningful model. Since $\hat{\theta}_{RST1}$ is not designed to protect against high leverage points this behavior was expected. Pregibon (1982)

offers a modified version, $\hat{\theta}_{RST2}$, to deal with leverage problems. Generally one can expect this modified estimator to behave more like our optimal estimators. One of the strengths of the bounded influence estimators is their ability to perform well in the presence of outliers as well as high leverage points.

Conclusions

Logistic regression is often performed on a routine basis in the analysis of binary data. The bounded influence procedures of this paper provide methods of fitting meaningful models in the presence of anomalous data. In addition these methods are amenable to asymptotic inference, a feature which is important whenever hypothesis testing is an objective of the analysis.

The procedures also supply useful tools for the data analyst interested in model building. Variable selection, as well as estimation, can be influenced by anomalous data; Pregibon (1982) cites such an example. Often the 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. This would have happened in the analysis of the food stamp data if models quadratic or piecewise linear in log(income + 1) were contemplated. In addition to variable selection the natural diagnostics from a robust fit, i.e. the weights and residuals, provide convenient methods of critically assessing the data and

model which serve as useful supplements to Pregibon's (1981) diagnostics. For example, with the food stamp data, an analyst, seeing the impact of case five, might question the validity of the observation or the appropriateness of the model over the full range of incomes.

A potentially significant shortcoming of the bounded influence procedures is their complexity and costliness in terms of computation. They do not exist for all choices of the bounding constant b and as the bound is lowered more computational finesse is required. In comparison Pregibon's estimators, which often produce similar models, are more easily computed using standard statistical packages. However, it is usually possible to obtain useful estimates with only moderate bounds on the influence and, in addition, our procedures admit an optimal asymptotic theory at the logistic model.

4. Proofs

In this section we prove the optimality results for $\hat{\theta}_{BL}$. The proofs of Theorem 1 and its corollary are, with the exception of notational complexities, nearly identical.

Proof of Theorem 2. Let X be a generic predictor having the empirical distribution $Pr(X = x_i) = 1/N$, $i = 1, \ldots, N$. Thus for example $\bar{x} = E_N(X)$ where E_N is the empirical expectation operator. Let $\psi = (y - F(x^T\theta))\ \omega(x, \theta)$ be any competitor of ψ_{BL}. Without loss of generality assume that $\psi = IC_\psi$, i.e. that ψ is in canonical form in the sense of Hampel (1978). This is equivalent to assuming that

$$E_N\ (F^{(1)}(X^T\theta)\ X\ \omega^T\ (X, \theta) = I_{p\times p}\ . \qquad (4.1)$$

Henceforth expressions within the expectation operator E_N are condensed by writing $F^{(1)}$ for $F^{(1)}(X^T\theta)$, ω for $\omega(X, \theta)$, etc. With this convention the following identity is easily established.

$$E_N(F^{(1)}(D_{BL}^{-1} X - \omega)\ (D_{BL}^{-1} X - \omega)^T) = E_N(F^{(1)}\ D_{BL}^{-1} XX^T\ D_{BL}^{-1}) \qquad (4.2)$$
$$- E_N(F^{(1)}\ D_{BL}^{-1} X\ \omega^T) - E_N\ (F^{(1)}\ D_{BL}^{-1}\ \omega\ X^T) + E_N\ (F^{(1)}\ \omega\ \omega^T)\ .$$

D_{BL} is the matrix defined by (2.2) corresponding to ψ_{BL}. As a consequence of (4.1) $\omega(\cdot,\cdot)$ persists on the right hand side of (4.2) only through the term $E_N\ (F^{(1)}\ \omega\ \omega^T)$, which is just V_ψ. Therefore trace($V_\psi\ V_{BL}^{-1}$) is, neglecting an additive constant independent of ψ, proportional to

$$E_N\ (F^{(1)}\ (D_{BL}^{-1} X - \omega)^T\ V_{BL}^{-1}\ (D_{BL}^{-1} X - \omega))\ . \qquad (4.3)$$

Let $\phi(X, \theta) = V_{BL}^{-\frac{1}{2}}\omega\ (X, \theta)$; (4.3) can then be written as

$$E_N(F^{(1)}\ \|\ \phi - V_{BL}^{-\frac{1}{2}}\ D_{BL}^{-1}\ X\ \|^2). \qquad (4.4)$$

Note that

$$\|\ \phi\ (x, \theta)\ \|^2 \leq\ b^2\ /\ m^2\ (x^T\ \theta)\quad \text{for all } x \qquad (4.5)$$

if and only if

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

Subject to (4.5) expression (4.4) is minimized as a functional of ϕ by taking

$$\phi(x, \theta) = V_{BL}^{-\frac{1}{2}}\ D_{BL}^{-1}\ x\ r^{\frac{1}{2}}\ (b^2\ /\ (m^2\ (x^T\ \theta)\ \|\ V_{BL}^{-\frac{1}{2}}\ D_{BL}^{-1}\ x\ \|^2)). \qquad (4.6)$$

Condition (4.1) insures that this is unique almost surely. Equations (2.2), (2.3), and (2.11) imply $D_{BL}^{-1} V_{BL}^{-1} D_{BL}^{-1} = Q^{-1}$ thus in terms of $\omega(\cdot,\cdot)$, (4.6) can be written as

$$\omega(x, \theta) = D_{BL}^{-1} x \, r^{\frac{1}{2}}(b^2 / (m^2 (x^T \theta) x^T Q^{-1} x))$$

proving the theorem.

Proof of Corollary 2.1. Again assume that all scores are in canonical form. Define

$$S = \{\psi \in \mathcal{M} : \sup_{(y,x)} \psi^T V_\psi^{-1} \psi \leq b^2 \} ;$$

$$S_{BL} = \{\psi \in \mathcal{M} : \sup_{(y,x)} \psi^T V_{BL}^{-1} \psi \leq b^2 \}.$$

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

$$\psi_{opt}^T V_{BL}^{-1} \psi_{opt} \leq \psi_{opt}^T V_{\psi_{opt}}^{-1} \psi_{opt} \leq b^2,$$

and hence ψ_{opt} is in S_{BL}. Set $\bar{S} = S \cap S_{BL}$. $\bar{S}$ is nonempty; it contains ψ_{BL} and ψ_{opt}. For any $\psi \in \bar{S}$ we known that $V_{\psi_{opt}} \leq V_\psi$ and hence

$$\text{trace}(V_{\psi_{opt}} V_{BL}^{-1}) \leq \text{trace}(V_\psi V_{BL}^{-1}) \quad \text{for all } \psi \text{ in } \bar{S} .$$

But Theorem 2 implies that ψ_{BL}, when defined, and in canonical form, is the almost everywhere unique minimizer of trace $(V_\psi V_{BL}^{-1})$ among all ψ in $\bar{S}$. The equivalence of ψ_{opt} and ψ_{BL} follows.

Acknowledgments

The research of the first two authors was supported by the Air Force Office of Scientific Research under grant AFOSR-80-0080, while that of the third author was supported by the national Science Foundation under grant MCS-81-00748.

References

Aranda-Ordaz, F. J. (1981), "On two families of transformations to additivity for binary response data," *Biometrika* 68, 357-63.

Berkson, J. (1951), "Why I prefer logists to probits." *Biometrics*, 7, 327-339.

Carroll, R. J., Spiegelman, C. H., Gordan Lan, K. K., Bailey, K. T., and Abbott, R. D. (1984), "On errors-in-variables for binary regression models." *Biometrika*, 71, 19-25.

Cook, D. R. and Weisberg, S. (1983), "Comment" on "Minimax Aspects of Bonded-Influence Regression." by Huber. *Journal of the American Stastistical Association*, 78, 74-75.

Cox, D. R. (1970), *Analysis of Binary Data*. London: Chapman and Hall Ltd.

Efron. B. (1975), "The efficiency of logistic regression compared to normal discriminant analysis." *Journal of the American Statistical Association*, 70, 892-898.

Guerrero, V. M. and Johnson, R. A. (1982), "Use of the Box-Cox transformation with binary response models." *Biometrika*, 69, 309-14.

Hampel, F. R. (1968), "Contributions to the theory of robust statistics," Ph.D. thesis, University of California, Berkeley.

_____ (1974), "The influence curve and its role in robust estimation." *Journal of the American Statistical Association*, 69, 383-394.

_____ (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.

Krasker, W. S. and Welsch, R. E. (1982), "Efficient bounded influence regression estimation using alternative definitions of sensitivity." *Journal of the American Statistical Association*, 77, 595-605.

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

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

Pregibon, D. (1981), "Logistic regression diagnostics." *Annals of Statistics*, 9, 705-724.

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

Ruppert, D. (1983), "On the bounded-influence regression estimator of Krasker and Welsch," (to appear in the *Journal of American Statistical Association*).

Stahel, W. A. (1981), "Robuste schaetzungen: Infinitesimale optimalitaet und schaetzungen von kovarianzmatrizen," Ph.D. thesis, Swiss Federal Institute of Technology, Zurich.

Stefanski, L. A. (1983), "Influence and measurement error in logistic regression." Institute of Statistics Mimeo Series No. 1548, University of North Carolina, Chapel Hill.

Table 1

Skin Vaso-constriction Data*

	$\hat{\theta}_{ML}$	$\hat{\theta}_{BI}(C=0)$ $b=3.7p^{\frac{1}{2}}$	$\hat{\theta}_{BL}$ $b=1.5LB$	$\hat{\theta}_{RST1}$
intercept	-2.92 (-2.27)	- 5.70 (-2.33)	-5.65 (-2.32)	-5.34 (-2.40)
log(rate)	4.63 (2.59)	8.08 (2.44)	7.96 (2.43)	7.62 (2.53)
log(volume)	5.22 (2.81)	9.12 (2.45)	8.91 (2.42)	8.60 (2.57)
observed efficiency	100%**	90%	87%	98%***

* Table entries include the parameter estimate, asymptotic z-statistics (in parentheses), and the measure of observed efficiency.

** By definition.

*** This efficiency is somewhat optimistic, see Section 3.

Table 2

Coronary Heart Disease Data *

	$\hat{\theta}_{ML}$	$\hat{\theta}_{BI}(C=0)$ $b=3.7p^{\frac{1}{2}}$	$\hat{\theta}_{BL}$ $b=1.25LB$	$\hat{\theta}_{RST1}$	$\hat{\theta}_{ML}$ sans 4,35
intercept	-46.70 (-2.96)	-55.47 (-2.99)	-49.55 (-2.95)	-55.24 (-2.93)	-76.92 (-3.09)
log(blood pressure)	4.24 (3.53)	4.79 (3.38)	4.09 (3.31)	5.27 (3.43)	5.48 (3.49)
log(cholesterol level)	4.82 (1.88)	5.95 (2.06)	5.46 (2.03)	5.49 (1.89)	9.27 (2.45)
observed efficiency	100%**	89%	91%	99%***	NA

* Table entries include the parameter estimate, asymptotic z-statistics (in parentheses), and the measure of observed efficiency.

** By definition.

*** This efficiency is somewhat optimistic, see Section 3.

Table 3

Food Stamp Data*

	$\hat{\theta}_{ML}$	$\hat{\theta}_{BI}(C=0)$ $b=3.5p^{\frac{1}{2}}$	$\hat{\theta}_{BL}$ $b=1.5LB$	$\hat{\theta}_{RST1}$	$\hat{\theta}_{ML}$ sans 5,66
intercept	0.93 (0.57)	4.26 (1.67)	4.14 (1.63)	1.02 (0.58)	6.88 (2.41)
tenancy	-1.85 (-3.46)	-1.85 (-3.43)	-1.81 (-3.39)	-2.28 (-3.51)	-2.02 (-3.53)
supplemental income	0.90 (1.79)	0.75 (1.46)	0.75 (1.46)	1.04 (1.90)	0.76 (1.41)
log(monthly income	-0.33 (-1.22)	-0.89 (-2.06)	-0.86 (-2.02)	-0.38 (-1.28)	-1.33 (-2.74)
observed efficiency	100%**	92%	90%	98%***	NA

* Table entries include the parameter estimate, asymptotic z-statistics (in parentheses), and the measure of observed efficiency.

** By definition.

*** This efficiency is somewhat optimistic, see Section 3.

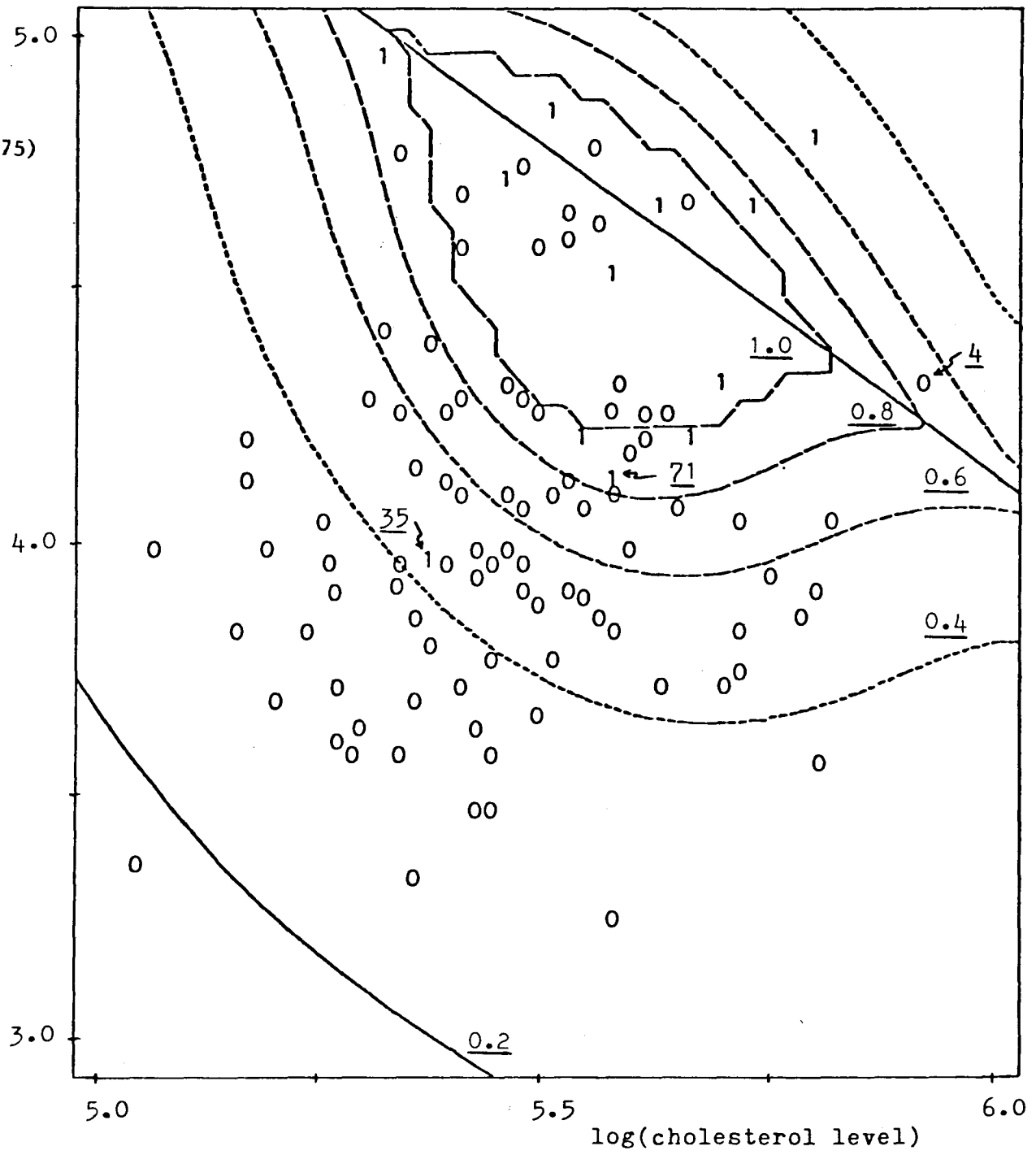

Figure 1. Plot of CHD Data. Cases of coronary heart disease are indicated by the numeral 1. Overlaying the data is a contour plot of the bounded leverage weights indicating potentially influential regions of the design space. Potential influence increases with decreasing weight. For reference the line corresponding to a 50% chance of heart disease is superimposed.

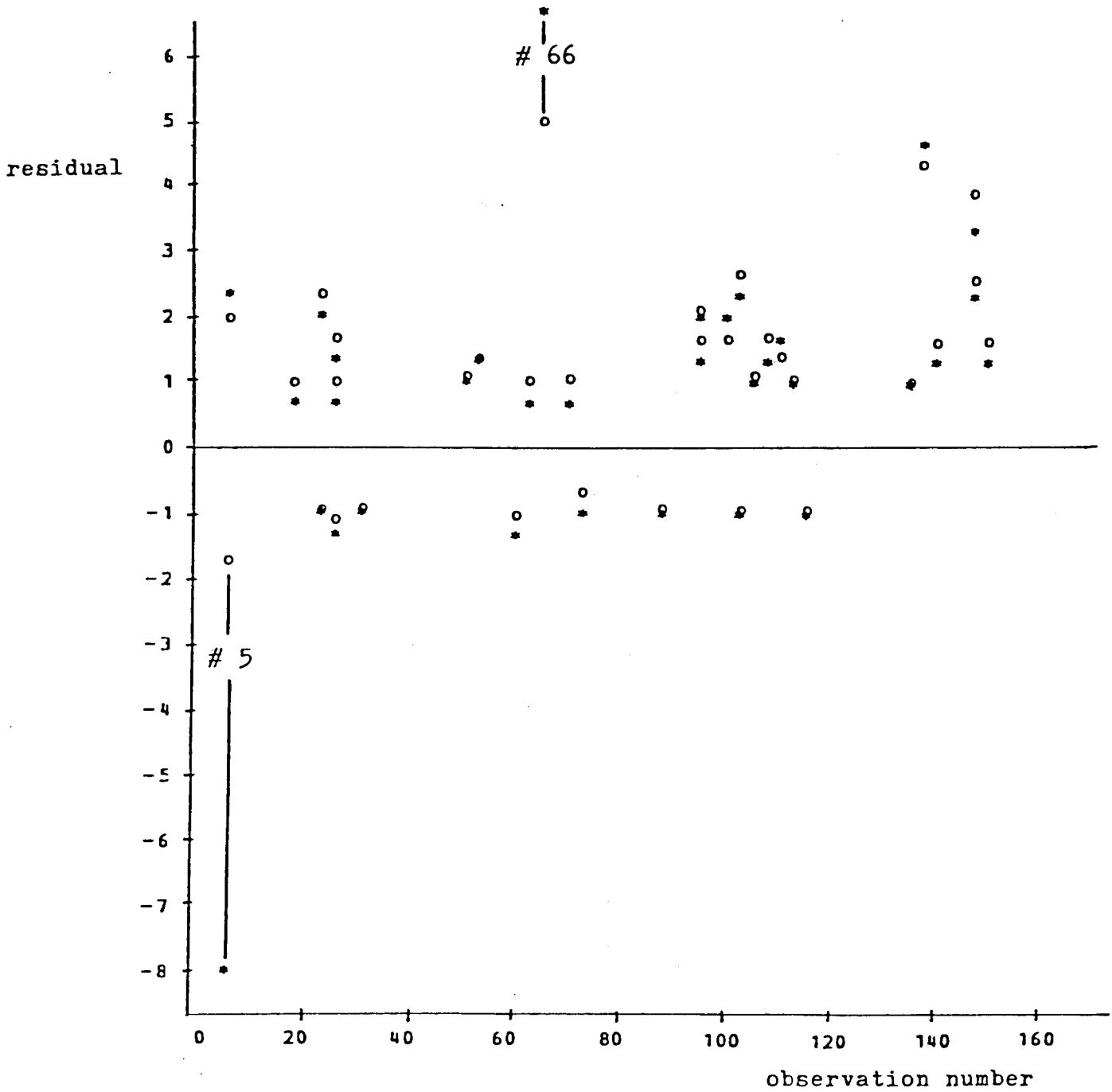

Figure 2. Residual Plots for FS Data. Maximum likelihood residuals are indicated by circles 'o'; residuals from the bounded leverage fit by asterisks '*'. For both estimation procedures residuals are defined as in Cox(1970), p. 96. Negligible residuals have been omitted for clarity.