The Annals of Statistics
1981, Vol. 9, No. 4, 705–724

LOGISTIC REGRESSION DIAGNOSTICS[1]

By Daryl Pregibon

Princeton University

A maximum likelihood fit of a logistic regression model (and other similar models) is extremely sensitive to outlying responses and extreme points in the design space. We develop diagnostic measures to aid the analyst in detecting such observations and in quantifying their effect on various aspects of the maximum likelihood fit. The elements of the fitting process which constitute the usual output (parameter estimates, standard errors, residuals, etc.) will be used for this purpose. With a properly designed computing package for fitting the usual maximum-likelihood model, the diagnostics are essentially "free for the asking." In particular, good data analysis for logistic regression models need not be expensive or time-consuming.

1. Introduction. Classically, logistic regression models were fit to data obtained under experimental conditions, for example, bioassay and related dose-response applications. The current use of logistic regression methods includes the analysis of data obtained in observational studies. In contrast to controlled experimentation, data from such studies can be notoriously "bad"—"bad" from the point of view of outlying responses (y), and "bad" from the point of view of extreme points in the design space ($\mathbf{X}$). The usual method of fitting logistic regression models, maximum likelihood, has good optimality properties in ideal settings, but is extremely sensitive to "bad" data of the above types.

For the normal-theory linear model, much is known about the effect on a least-squares fit of outlying responses and extreme design points, the latter frequently called "high-leverage points." An overview of diagnostic measures used in the analysis of linear models is discussed in Section 2.

An important extension of these diagnostic approaches is to nonlinear regression models, where presumably the effects of outliers and leverage points could be worse. This paper proposes diagnostic measures which should accompany the "usual" output from a maximum likelihood fit of a logistic regression model. As this model is a member of the class of generalized linear models (Nelder and Wedderburn, 1972), the methods described here were developed with the aim of being applicable to the entire class (see Pregibon, 1979). The general pattern of the computations could well lead to applications in models outside this specialized framework, say to the areas of time series analysis and survival models. The author is currently investigating the possibilities in the latter area.

In Section 3, we introduce the model, the relevant notation, and an example. In Section 4, we develop the basic building blocks of inexpensive regression diagnostics. In Section 5, we introduce a device to perturb the maximum likelihood fit and, accordingly, allow for suitable regression diagnostics to be deduced. The diagnostics are described in detail in Section 6. We close with a brief comment on various extensions and computational considerations.

2. An overview of regression diagnostics for the standard linear model. The standard linear model is specified as

$$y_i = \mathbf{x}_i\boldsymbol{\beta} + \epsilon_i, \qquad i = 1, \ldots, N$$

Received December, 1979; revised October, 1980.

[1] This work originated while the author was attending the University of Toronto, Canada, under support of a National Research Council of Canada Postgraduate Scholarship.

AMS 1970 *subject classifications*. Primary 62F99, 62J05; secondary 62P10.

Key words and phrases: Logistic regression, generalized linear models, regression diagnostics, residual analysis, iteratively reweighted least squares, leverage points, influence curves.

where $\mathbf{x}\boldsymbol{\beta} = x_1\beta_1 + \ldots + x_m\beta_m$ and $\epsilon \sim N(0, \sigma^2)$. The least-squares estimate $\hat{\boldsymbol{\beta}}$ is obtained by solving the normal equations

$$\mathbf{X}^T\mathbf{r} = \mathbf{X}^T(\mathbf{y} - \hat{\mathbf{y}}) = \mathbf{0}$$

where $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$ is the vector of fitted values. The solution of this linear system is $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$, which is sensitive to poorly-fit observations and extreme design points.

Presently, there is a fairly large battery of diagnostics available for detecting which observations exert undue influence on $\hat{\boldsymbol{\beta}}$. The two basic quantities that are most useful for this purpose are the residuals, $r_i = y_i - \mathbf{x}_i\hat{\boldsymbol{\beta}}$, and the projection matrix

$$\mathbf{M} = \mathbf{I} - \mathbf{H} = \mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T.$$

Essentially, the vector $\mathbf{r}$ describes the deviation of the observed data from the fit, and $\mathbf{M}$ the subspace in which $\mathbf{r}$ lies.

The residual vector $\mathbf{r}$ is important for the detection of ill-fitting points, but will not adequately point to observations which unduly influence the fit. In particular, large residuals are seldom associated with high-leverage points, whereas small residuals (which usually pass our inspection unnoticed) are typically of the opposite character.

The diagonal elements of $\mathbf{M}$ can direct us to such points (see for example, Hoaglin and Welsch, 1978). Influential points will tend to have small values of m_{ii}, much smaller than the average value $1 - m/N$. Hoaglin and Welsch (1978) suggest using $m_{ii} \leq 1 - 2m/N$ as a rough guide for determining whether a point is influential or not.

The quantities r_i and m_{ii} (or h_{ii}) are useful for detecting extreme points, but not for assessing their impact on the various aspects of the fit, e.g., parameter estimates, fitted values, goodness-of-fit measures, etc. Two approaches which attempt to quantify the effect of individual observations on the fit have been investigated:

(1) assessment by deletion (references [1, 2, 4, 5, 6, 11, 13]), and

(2) assessment by infinitesimal perturbations (references [2, 11, 13]).

In the former approach, one computes the change in some aspect of the fit incurred by deleting one or more data points. For single deletions (say the lth observation), the basic formula for the change in the least-squares estimate $\hat{\boldsymbol{\beta}}$ is given by

$$\Delta_l\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}(1) - \hat{\boldsymbol{\beta}}(0) = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_l r_l/m_{ll},$$

where (1) and (0) denote the presence and absence of the lth point in computing the quantity immediately to the left of $(\cdot)$.

A scalar measure summarizing the effect of deleting the lth observation over all coefficients is

$$c_l = (\Delta_l\hat{\boldsymbol{\beta}})^T\mathbf{X}^T\mathbf{X}(\Delta_l\hat{\boldsymbol{\beta}}) = r_l^2 h_{ll}/(1 - h_{ll})^2.$$

This measure, when appropriately standardized, can be given several interpretations (see Cook, 1977, 1979). One such interpretation is the confidence region displacement due to deleting the lth observation. Another is the sum of squared distances of $\Delta_l\hat{\mathbf{y}}$. In either case, individual observations which produce large values of c_l unduly influence the overall fit of the model. In many cases these observations are of interest in themselves, indicating a set of experimental conditions much different from the others.

Another type of summary diagnostic is the change in the residual sum of squares (RSS) due to deleting the lth observation:

$$\Delta_l\text{RSS} = \text{RSS}(1) - \text{RSS}(0) = r_l^2/m_{ll}.$$

This diagnostic provides much of the motivation for computing and plotting studentized residuals.

Note that all the above diagnostics can be readily computed for each observation as all the necessary bits (the basic building blocks) are available following the usual fit. Other diagnostics, too numerous to mention here, are almost exclusively functions of the residuals and the diagonal elements of $\mathbf{M}$. The off-diagonal elements of $\mathbf{M}$ will play an important role when we consider the joint effects of several observations on the fit.

The infinitesimal perturbation approach is obtained by specifying $\epsilon_i \sim N(0, \sigma^2/w_i)$ where

$$w_i = \begin{cases} w & i = l \\ 1 & \text{otherwise} \end{cases}$$

and $0 \le w \le 1$. According to this specification, the normal equations are modified as $\mathbf{X}^T\mathbf{W}\mathbf{r} = \mathbf{0}$ with $\mathbf{W} = diag\{w_i\}$, leading to the modified least squares estimate $\hat{\boldsymbol{\beta}}(w) = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{y}$. This estimate is related to $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}(1)$ via the basic formula

$$\hat{\boldsymbol{\beta}}(1) - \hat{\boldsymbol{\beta}}(w) = \frac{(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_l(1 - w)r_l}{\{1 - (1 - w)h_{ll}\}}$$

which simplifies to $\Delta_l\hat{\boldsymbol{\beta}}$ at $w = 0$. The effect of infinitesimal perturbations of the variance of the lth data point is easily obtained by differentiation of $\hat{\boldsymbol{\beta}}(w)$ leading to

$$\dot{\hat{\boldsymbol{\beta}}}(w) = \frac{\partial}{\partial w}\hat{\boldsymbol{\beta}}(w) = \frac{(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_l r_l}{\{1 - (1 - w)h_{ll}\}^2}.$$

In particular,

$$\dot{\hat{\boldsymbol{\beta}}}(1) = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_l r_l$$

$$\dot{\hat{\boldsymbol{\beta}}}(0) = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_l r_l/(1 - h_{ll})^2.$$

Evaluation at $w = 1$ describes local changes in $\hat{\boldsymbol{\beta}}(w)$ at the usual least squares solution. In the literature of robust and resistant estimation, this function is termed the influence curve of the estimate $\hat{\boldsymbol{\beta}}$. For diagnostic purposes, it tends to be conservative because an extreme observation tends to pull the fit toward itself, and infinitesimal changes at this point may not perturb the fit sufficiently to cause large changes in the coefficients (see, for example, Figure 3). On the other hand, evaluation at $w = 0$ describes the changes in $\hat{\boldsymbol{\beta}}(w)$ at the least squares fit to the data without the lth observation. This function can also be interpreted as an influence curve, but one which is sensitive to both outlying responses and extreme design points. In particular, note that the denominator of $\dot{\hat{\boldsymbol{\beta}}}(0)$ *is* $m_{ll}^2 = (1 - h_{ll})^2$, and small values of m_{ll} indicate extreme design points. This latter influence curve is more desirable for diagnostic purposes, although it has seen very little use (as yet) in the literature.

Since $\hat{\boldsymbol{\beta}}(w)$ is everywhere differentiable on the unit interval, application of the mean value theorem yields

$$\frac{\hat{\boldsymbol{\beta}}(1) - \hat{\boldsymbol{\beta}}(0)}{1 - 0} = \dot{\hat{\boldsymbol{\beta}}}(\bar{w}),$$

for some $\bar{w}$, $0 < \bar{w} < 1$. That is, the change in the least squares estimate of $\boldsymbol{\beta}$ due to dropping the lth observation is given by the derivative of $\hat{\boldsymbol{\beta}}(w)$ evaluated at an interior point of the unit interval. Accordingly, some writers (e.g. Cook and Weisberg, 1979) have termed $\Delta_l\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}(1) - \hat{\boldsymbol{\beta}}(0)$ as the empirical influence curve of the estimate.

All three versions of the influence curve provide differential information concerning the effect of individual observations on the fit. Since $\Delta_l\hat{\boldsymbol{\beta}} = \dot{\hat{\boldsymbol{\beta}}}(\bar{w})$ for some $\bar{w}$ in the unit interval, it tends to be a reasonable compromise between conservatism (at $w = 1$) and liberalism (at $w = 0$). Apart from the intuitive appeal, it is perhaps this reason that most regression diagnostics are formulated by the deletion method. As our experience grows with understanding and interpreting infinitesimal perturbations at $w = 0$, this may no longer remain the case.

Summary type diagnostics can be formed for the infinitesimal perturbation model in an analogous fashion as in the deletion model. The interested reader is referred to Pregibon (1979).

3. Background and notation for the logistic regression model.

3.1. *The Unstructured Case.* Consider a single binomial response $y \sim B(n, p)$. If we let $\theta = \text{logit}(p) = \log\{p/(1-p)\}$, the probability function of y can be written as

$$f(y; \theta) = \exp\{y\theta - a(\theta) + b(y)\} \qquad y = 0, 1, \ldots, n$$

with $a(\theta) = n \log(1 + e^{\theta})$, $b(y) = \log\binom{n}{y}$ and where throughout this paper $\log(\cdot) = \log_e(\cdot)$. Up to an arbitrary constant, the logarithm of $f(y; \theta)$,

$$l(\theta; y) = y\theta - a(\theta) + b(y),$$

is the loglikelihood function of θ. The score and information functions are given by

$$s(\theta; y) = \frac{\partial}{\partial\theta} l(\theta; y) = y - \dot{a}(\theta) = y - np$$

$$v(\theta; y) = -\frac{\partial}{\partial\theta} s(\theta; y) = \ddot{a}(\theta) = np(1-p),$$

where a with k dots above it denotes $(\partial^k/\partial\theta^k)a(\theta)$. Standard results yield $\mathbf{E}\{s(\theta; y)\} = 0$ (or $\mathbf{E}(y) = np = \dot{a}(\theta)$) and $\text{Var}(y) = np(1-p) = \ddot{a}(\theta)$. Also, since $s(\hat{\theta}; y) = 0$ at the maximum likelihood estimate (m.l.e.) $\hat{\theta}$, we have $\hat{\theta} = \dot{a}^{-1}(y) = \text{logit}(y/n)$ as the m.l.e. of θ based on a single binomial observation y.

Given a sample of N independent binomial responses $y_i \sim B(n_i, p_i)$, the loglikelihood function for the sample is the sum of individual loglikelihood contributions:

$$l(\boldsymbol{\theta}; \mathbf{y}) = \textstyle\sum_{i=1}^{N} l(\theta_i; y_i) = \sum_{i=1}^{N} \{y_i\theta_i - a(\theta_i) + b(y_i)\}.$$

3.2. *The Logistic Regression Model.* The likelihood function $l(\boldsymbol{\theta}; \mathbf{y})$ is over-specified—there are as many parameters as observations. Given a set of m explanatory variables $\{X_1, X_2, \ldots, X_m\}$, the logistic regression model utilizes the relationship

$$\boldsymbol{\theta} = \text{logit}(\mathbf{p}) = \mathbf{X}\boldsymbol{\beta}$$

as the description of the systematic component of the response $\mathbf{y}$. In terms of the m-dimensional parameter $\boldsymbol{\beta}$, we have the loglikelihood function:

$$(1) \qquad l(\mathbf{X}\boldsymbol{\beta}; \mathbf{y}) = \textstyle\sum_{i=1}^{N} l(\mathbf{x}_i\boldsymbol{\beta}; y_i) = \sum_{i=1}^{N} y_i\mathbf{x}_i\boldsymbol{\beta} - a(\mathbf{x}_i\boldsymbol{\beta}) + b(y_i).$$

The m.l.e. maximizes (1) and is a solution (assumed unique) to $(\partial/\partial\hat{\boldsymbol{\beta}})l(\mathbf{X}\hat{\boldsymbol{\beta}}; \mathbf{y}) = \mathbf{0}$. In particular, $\hat{\boldsymbol{\beta}}$ satisfies the system of equations:

$$\textstyle\sum_{i=1}^{N} x_{ij}(y_i - \dot{a}(\mathbf{x}_i\hat{\boldsymbol{\beta}})) = 0 \qquad j = 1, \ldots, m.$$

Writing $\mathbf{s} = \mathbf{y} - \dot{a}(\mathbf{X}\hat{\boldsymbol{\beta}}) = \mathbf{y} - \mathbf{n}\hat{\mathbf{p}}$, the matrix formulation of the likelihood equations is

$$\mathbf{X}^T\mathbf{s} = \mathbf{X}^T(\mathbf{y} - \hat{\mathbf{y}}) = \mathbf{0}.$$

These equations, although very similar to their normal theory counterparts, are nonlinear in $\hat{\boldsymbol{\beta}}$, and iterative methods are required to solve them. Typically, when second derivatives are easy to compute (in the present case $-(\partial/\partial\hat{\boldsymbol{\beta}})\mathbf{X}^T\mathbf{s} = \mathbf{X}^T\mathbf{V}\mathbf{X}$ with $\mathbf{V} = \text{diag}\{\ddot{a}(\mathbf{x}_i\hat{\boldsymbol{\beta}})\}$), the Newton-Raphson method is employed. This leads to the iterative scheme

$$\boldsymbol{\beta}^{t+1} = \boldsymbol{\beta}^t + (\mathbf{X}^T\mathbf{V}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{s} \qquad t = 0, 1, \ldots *$$

where both $\mathbf{V}$ and $\mathbf{s}$ are evaluated at $\boldsymbol{\beta}^t$. At convergence ($t = *$), we take $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}^*$, and denote the fitted values $n_i\hat{p}_i$ by $\hat{y}_i$. The estimated variance of y_i is $v_{ii} = n_i\hat{p}_i(1 - \hat{p}_i)$.

A most useful way to view the iterative process outlined above is by the method of iteratively reweighted least-squares (IRLS). This is obtained by employing the pseudoobservation vector $\mathbf{z}^t = \mathbf{X}\boldsymbol{\beta}^t + \mathbf{V}^{-1}\mathbf{s}$, upon which the above equation becomes

$$\boldsymbol{\beta}^{t+1} = (\mathbf{X}^T\mathbf{V}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}\mathbf{z}^t.$$

At convergence, we have $\mathbf{z} = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{V}^{-1}\mathbf{s}$. Thus we may write the m.l.e. of $\boldsymbol{\beta}$ as $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{V}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}\mathbf{z}$. This form of the estimate will provide the basis of extending the results of Section 2 to the logistic regression case.

3.3. *The Standard Output from a Maximum Likelihood Fit.* Once the model has been fitted (that is, we have the m.l.e. $\hat{\boldsymbol{\beta}}$), various quantities from the fitting process are available to the analyst. Typically, these quantities consist of a subset of the following:

(a) the estimated parameter vector, $\hat{\boldsymbol{\beta}}$;
(b) the individual coefficient standard errors, s.e.$(\hat{\boldsymbol{\beta}}_j)$;
(c) the estimated covariance matrix of $\hat{\boldsymbol{\beta}}$, $\mathrm{Var}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}^T\mathbf{V}\mathbf{X})^{-1}$;
(d) the chi-squared goodness-of-fit statistic $\chi^2 = \sum_{i=1}^N s_i^2/v_{ii}$;
(e) the individual components of χ^2, namely $\chi_i = s_i/\sqrt{v_{ii}} = (y_i - n_i\hat{p}_i)/\sqrt{n_i\hat{p}_i(1 - \hat{p})}$;
(f) the deviance $D = -2\{l(\mathbf{X}\hat{\boldsymbol{\beta}}; \mathbf{y}) - l(\hat{\boldsymbol{\theta}}; \mathbf{y})\}$, where $l(\hat{\boldsymbol{\theta}}; \mathbf{y})$ refers to the maximum of the loglikelihood function based on fitting each point exactly, i.e., $\theta_i = \mathrm{logit}(y_i/n_i)$.

Asymptotic arguments suggest that the deviance and chi-squared statistics have the same limiting null $\chi^2(N - m)$ distribution, and hence provide some measure of the appropriateness of the fitted model.

As an illustration of the standard output from a maximum likelihood fit, we introduce an example from Finney (1947). The data, listed in Table 1, were obtained in a carefully controlled study of the effect of the rate and volume of air inspired on a transient vaso-constriction in the skin of the digits. The nature of the measurement process was such that only the occurrence or nonoccurrence of vaso-constriction could be reliably measured. Three subjects were involved in the study: the first contributed 9 responses, the second contributed 8 responses, and the third contributed 22 responses. A plot of the data appears

TABLE 1
Listing of Finney's data on vaso-constriction in the skin of the digits. The binary response y indicates the occurrence (1) *or nonoccurrence* (0) *of vaso-constriction.*

Volume	Rate	Response	Volume	Rate	Response
3.7	.825	1	1.8	1.8	1
3.5	1.09	1	.4	2	0
1.25	2.5	1	.95	1.36	0
.75	1.5	1	1.35	1.35	0
.8	3.2	1	1.5	1.36	0
.7	3.5	1	1.6	1.78	1
.6	.75	0	.6	1.5	0
1.1	1.7	0	1.8	1.5	1
.9	.75	0	.95	1.9	0
.9	.45	0	1.9	.95	1
.8	.57	0	1.6	.4	0
.55	2.75	0	2.7	.75	1
.6	3.	0	2.35	.03	0
1.4	2.33	1	1.1	1.83	0
.75	3.75	1	1.1	2.2	1
2.3	1.64	1	1.2	2.0	1
3.2	1.6	1	.8	3.33	1
.85	1.415	1	.95	1.9	0
1.7	1.06	0	.75	1.9	0
			1.3	1.625	1

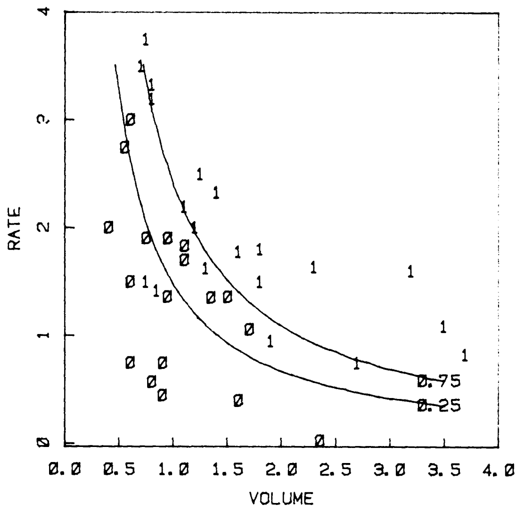

FIG. 1. *Scatter plot of Finney's data:* (0) *and* (1) *represent the nonoccurrence and occurrence of vaso-constriction of the skin. The curves labeled* .25 *and* .75 *are the* 25% *and* 75% *contours for the logit fit.*

in Figure 1. The two lines superimposed on the scatter plot are the estimated 25% and 75% contours corresponding to the model:

$$\text{logit}(p) = \beta_1 + \beta_2 \log(\text{RATE}) + \beta_3 \log(\text{VOLUME}).$$

The estimated coefficients and their standard errors are

$$\hat{\beta}_1 = -2.875\ (1.319)$$

$$\hat{\beta}_2 = 5.179\ (1.862)$$

$$\hat{\beta}_3 = 4.562\ (1.835).$$

The deviance for the fit is 29.23 on 36 degrees of freedom, and the corresponding chi-squared statistic is 34.15. Both are less than their asymptotic expectation of 36, indicating no gross inadequacies with the model. The ordered components of χ^2 are plotted against standard normal quantiles in Figure 2. Evidently, two observations, the 4th and 18th, are not well fit by the model—their χ_i residuals deviate from the straight-line configuration of the others. These two points correspond to the only positive responses outside the 25% contour (see Figure 1). As these observations are not really associated with extreme values in the design space, their effect on the fit might presumably be small. From the information thus far presented, we have no reason to believe otherwise.

4. The basic building blocks of regression diagnostics.

4.1. *Preliminaries.* After fitting a logistic regression model, and prior to drawing inferences from it, the natural succeeding step is that of critically assessing the fit. In practice however, this assessment is rarely considered, and seldom carried out. The basic reasons are

(i) the lack of routine methods (in the literature and in standard computing packages) for performing such an analysis, and

(ii) the presumably high costs (in analyst and computer time) of doing so.

The role of a regression diagnostician is to provide routine methods of model sensitivity analysis which are both intuitively appealing and inexpensive. Clearly this requires a

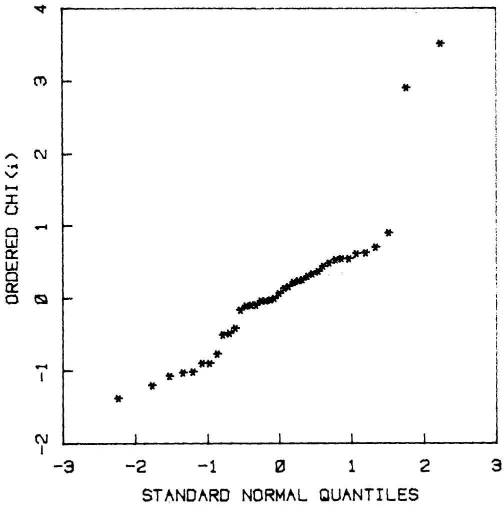

FIG. 2. *Gaussian probability plot of Finney's data: the ordered components of* χ^2 *vs standard normal quantiles.*

thorough understanding of the model *and* the nature of the fitting process (here, maximum likelihood).

In what follows, we assume that we have fit a logistic regression model by maximum likelihood, and that we have the components of that fit. We want to derive useful and informative diagnostic measures from these components to supplement the "usual" output. These measures should readily identify observations that are not well explained by the model, as well as those dominating some important aspect of the fit. In some cases, this analysis may reveal systematic departures of the data from the model, though, in general, this is not to be expected. A routine method for detecting systematic departures is given by Pregibon (1980; Chapter 3, 1979).

4.2. *The Basic Building Blocks.* For the logistic regression model, the basic building blocks for the identification of outlying and influential points will again be a residual vector and a projection matrix. For the linear model, residuals are rather uniquely defined (apart from standardization), whereas for the logistic regression model, residuals can be defined on several (at least three) scales. The two which we find most useful are the components of chi-squared, given in 3.3(e), and the components of deviance, $D = \sum d_i^2$:

$$d_i = \pm \sqrt{2}\{l(\hat{\theta}_i; y_i) - l(\mathbf{x}_i\hat{\boldsymbol{\beta}}; y_i)\}^{1/2},$$

where the plus or minus is used according as $\hat{\theta}_i > \mathbf{x}_i\hat{\boldsymbol{\beta}}$ or $\hat{\theta}_i < \mathbf{x}_i\hat{\boldsymbol{\beta}}$. Note that d_i is defined for all values of y_i even though $\hat{\theta}_i$ may not be. In particular, at $y = 0$, $d^2 = -2n \log(1 - \hat{p})$ and at $y = n$, $d^2 = -2n \log(\hat{p})$. Both χ^2 and D are measures of the goodness-of-fit of the model. The former measures the relative deviations between the observed and fitted values, whereas the latter measures the disagreement between the maxima of the observed and fitted loglikelihood functions. In either case, large individual components indicate observations poorly accounted for by the model.

The analog of the projection matrix for the logistic model will also be denoted by **M**, which in its general form is given as

$$\mathbf{M} = \mathbf{I} - \mathbf{H} = \mathbf{I} - \mathbf{V}^{1/2}\mathbf{X}(\mathbf{X}^T\mathbf{V}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{1/2}.$$

The usefulness of **M** arises as a consequence of the IRLS formulation described earlier. In particular, as $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{V}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}\mathbf{z}$, the vector of pseudo-residuals is given by

$$\mathbf{z} - \mathbf{X}\hat{\boldsymbol{\beta}} = \{\mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{V}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}\}\mathbf{z} = \mathbf{V}^{-1/2}\mathbf{M}\mathbf{V}^{1/2}\mathbf{z}.$$

Using the fact that $\mathbf{z} = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{V}^{-1}\mathbf{s}$, this can be written as $\mathbf{V}^{-1}\mathbf{s} = \mathbf{V}^{-1/2}\mathbf{M}\mathbf{V}^{-1/2}\mathbf{s}$. Premultiplication by the diagonal matrix $\mathbf{V}^{1/2}$ yields $\chi = \mathbf{M}\chi$, where $\chi = \mathbf{V}^{-1/2}\mathbf{s}$. Thus, as in the linear model case, $\mathbf{M}$ is symmetric, idempotent, and spans the residual (χ) space. This suggests that small m_{ii} should be useful in detecting extreme points in the design space. Experience has shown this to be the case.

The hypothetical data set listed in Table 2 and plotted in Figure 3 gives evidence supporting this contention. There is one potentially influential point, the last one in the list. The figure also displays the maximum likelihood fit of a straight line to the data (with and without the last point). The differences in fit between the two are striking. The values of m_{ii} are given in the last column of Table 2. The influential point has the smallest value of m_{ii}; this will always be the case whenever a point is far removed from the others in the design space.

This simple example is also useful in pointing out that the IRLS analogy can be carried only so far. In particular, as $v_{ii} = n_i\hat{p}_i(1 - \hat{p}_i)$ (which for fixed $\hat{\beta}$ decreases toward zero as $\mathbf{x}_i$ increases) is interpreted as the weight associated with the ith observation, it would seem that the logistic regression model is protected against extreme design points as they are automatically downweighted in the fitting procedure. However, the fact of the matter is that the weights are an artifact of the IRLS formulation and have nothing to do with the likelihood equations $\mathbf{X}^T\mathbf{s} = \mathbf{0}$ which determine $\hat{\boldsymbol{\beta}}$. That is, in the logistic regression model, *the weights v_{ii} are determined by the fit,* which should not be confused with a true weighted least-squares problem where *the weights determine the fit.* For those readers who remain doubtful of this subtle difference in interpretation of the weights, we suggest moving the last point in Figure 3 further and further to the right and observing the consequences.

In most cases, the examination of χ_i, d_i and m_{ii} will call attention to outlying and influential points. In some cases, combinations of these (for example, studentized residuals) will also be useful. For displaying these quantities, index plots are generally (and, if the order of the observations is important, *strongly*) suggested: that is, plots of χ_i vs i, d_i vs i and m_{ii} vs i. In particular cases, plots of these building blocks against the fitted values could prove useful.

Another useful plot is derived from the matrix

$$\mathbf{H}^* = \mathbf{V}^{1/2}\mathbf{X}^*(\mathbf{X}^{*T}\mathbf{V}\mathbf{X}^*)^{-1}\mathbf{X}^{*T}\mathbf{V}^{1/2}, \qquad \mathbf{X}^* = (\mathbf{X}; \mathbf{z}).$$

A little algebra gives $h_{ii}^* = h_{ii} + \chi_i^2/\chi^2$ as a diagonal element with $0 \le h_{ii}^* \le 1$ and ave(h_{ii}^*)

TABLE 2
Listing of an hypothetical data set and the basic building blocks associated with the maximum likelihood fit of a logistic regression model.

Data		Building Blocks		
x_i	logit (y_i/n_i)	χ_i	d_i	m_{ii}
1	0.5	−0.506	−0.496	0.7192
2	0.5	−0.615	−0.600	0.7885
3	0.8	−0.247	−0.244	0.8409
4	1.0	−0.055	−0.054	0.8773
5	1.3	0.248	0.252	0.8989
6	1.3	0.157	0.159	0.9070
7	1.6	0.417	0.431	0.9030
8	1.8	0.535	0.561	0.8882
9	2.1	0.715	0.768	0.8640
10	2.1	0.642	0.686	0.8316
17	1.0	−1.383	−1.227	0.4813

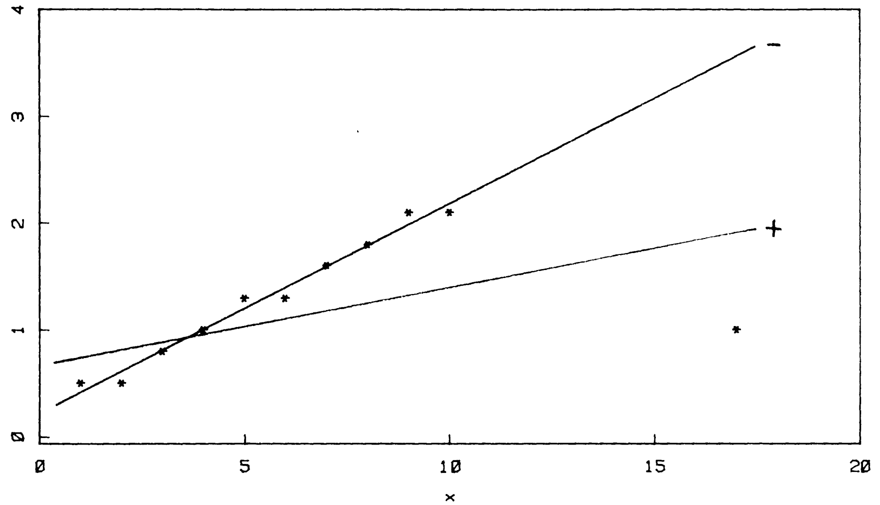

FIG. 3. *Scatter plot of an hypothetical data set and two fitted logit models. The labelling of the lines indicate whether the last point was (+) or wasn't (−) included in the fit.*

$= (m + 1)/N$. Thus, values of h_{ii}^* near unity correspond to observations which are poorly-fit (large relative χ_i^2), extreme in the design space (large h_{ii}), or both. A scatter plot of χ_i^2/χ^2 vs h_{ii} will display these characteristics. For calibration purposes, contours of constant h_{ii}^* (lines with slope -1) can be superimposed on the scatter plot to clearly expose large components of h_{ii}^*.

Returning to the binary logistic regression example, Figure 4 displays the index plots of m_{ii}, χ_i, and d_i based on the fitted logit model. At a glance, it is clear that the 4th and 18th observations are not well fit by the model. The 31st observation has the smallest value of m_{ii}. Reference to Figure 1 indicates that it is the right-most positive response between the 25% and 75% contours. Even though it is less than the rough cutoff of $.8466 = 1 - (2 \times 3)/39$, we will later see that its effect on the fit is minor in comparison to the previously identified points. Figure 5 displays the scatter plot of χ_i^2/χ^2 vs h_{ii} which effectively summarizes the information provided by the index plots.

5. Model perturbations and one-step estimates. The quantities introduced in the previous section should indicate which (if any) of the observations (1) are not well explained by the model, or (2) are dominating some aspect of the fit. The quantities, however, cannot adequately measure the effect on the many components of the fitted model. In this section we attempt to refine our understanding of each observation's effect on the fitted model. We proceed by first introducing a simple method of perturbing our usual model; this facilitates the study of the effects of individual points. These methods are generalizations of the ones used by Welsch and Kuh (1977) for the normal-theory linear model.

Consider first an unstructured set of binomial data $\{y_i: i = 1, \ldots, N\}$, the loglikelihood of which may be written as the sum of N individual contributions, $l(\theta; y_i)$. Alternatively, if there are K distinct values, with the multiplicity of each given by the observation count w_k, the loglikelihood function may be written as

$$l(\theta; \mathbf{y}) = \sum_{k=1}^{K} w_k l(\theta; y_k) = \sum_{k=1}^{K} w_k\{y_k\theta - a(\theta) + b(y_k)\}.$$

When we consider structured observational data, seldom will there be multiplicities greater than unity since all components of the observations vector $\{y_i, x_{i1}, \ldots, x_{im}\}$ must match for this to occur. If all multiplicites are unity, $K = N$, $k = i$, and the loglikelihood function

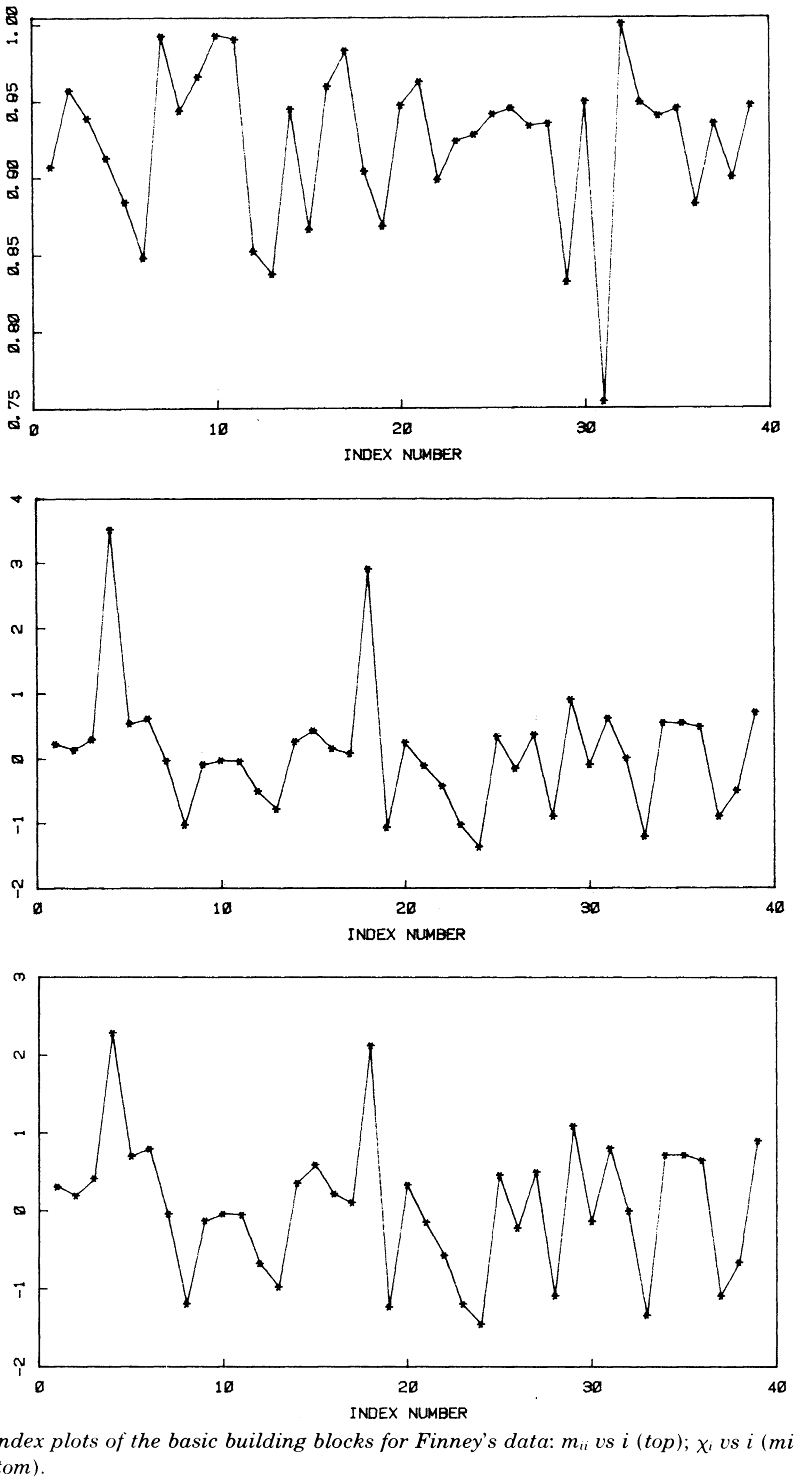

FIG. 4. *Index plots of the basic building blocks for Finney's data*: m_{ii} *vs i* (*top*); χ_i *vs i* (*middle*); d_i *vs i* (*bottom*).

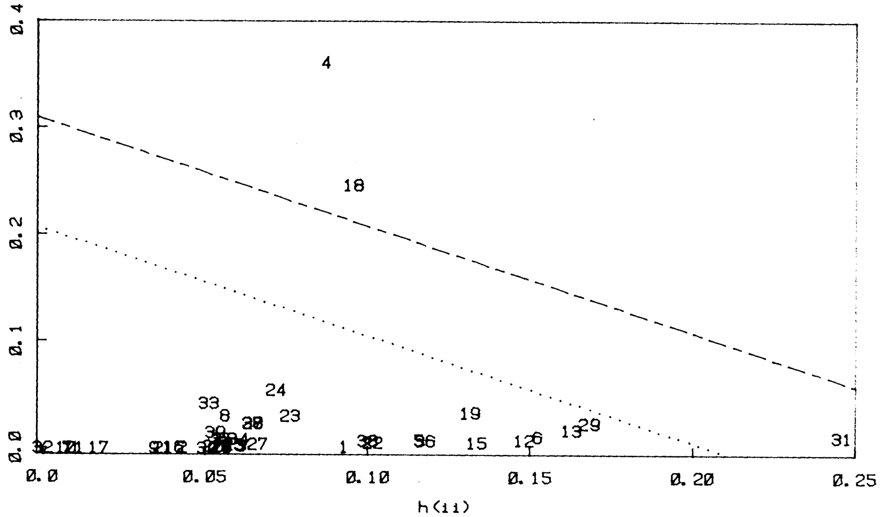

FIG. 5. *Scatter plot of the basic building blocks for Finney's data*: χ_i^2/χ^2 *vs* h_{ii}. *The dashed lines correspond to the* $2 \times \text{ave}(h_{ii}^*)$ *and* $3 \times \text{ave}(h_{ii}^*)$.

may be expressed as

$$l_w(\mathbf{X}\boldsymbol{\beta};\, \mathbf{y}) = \textstyle\sum_{i=1}^{N} w_i l(\mathbf{x}_i\boldsymbol{\beta};\, y_i), \tag{2}$$

where $w_i = 1$ for all i. The reason for leaving the observation count in the loglikelihood specification is that it permits simple perturbations of the model attributable to changes in individual points. This allows us to study the effect of each observation on important aspects of the fit.

To begin this study, consider

$$w_i = \begin{cases} w & \text{for } i = l \\ 1 & \text{otherwise} \end{cases}$$

with $0 \le w \le 1$. The maximum likelihood estimate of $\boldsymbol{\beta}$ will now be a function of w and can be obtained by maximizing (2). This is equivalent to solving the system of equations:

$$\textstyle\sum_{i=1}^{N} x_{ij} w_i s_i = 0 \qquad j = 1, \ldots, m.$$

In matrix form these equations are $\mathbf{X}^T\mathbf{W}\mathbf{s} = \mathbf{0}$. The matrix of mixed partial derivatives of l_w with respect to β_j and β_k is $-\mathbf{X}^T\mathbf{V}^{1/2}\mathbf{W}\mathbf{V}^{1/2}\mathbf{X}$. Thus, the Newton-Raphson method leads to the sequence of estimates

$$\boldsymbol{\beta}^{t+1}(w) = \boldsymbol{\beta}^{t}(w) + (\mathbf{X}^T\mathbf{V}^{1/2}\mathbf{W}\mathbf{V}^{1/2}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{s}. \tag{3}$$

At $w = 1$, $\mathbf{W}$ is the identity matrix, and the above simplifies to the usual maximum likelihood iterative procedure for $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}(1)$.

Our objective is to determine individual effects with a minimal effort. For big effects, it is enough to know their direction, whereas for small changes, more precision is required. Useful results can be obtained for this purpose by starting from the usual maximum likelihood fit (that is, $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}(1)$) and terminating the sequence (3) after one step.[2] In this case, equation (3) becomes

$$\hat{\boldsymbol{\beta}}^1(w) = (\mathbf{X}^T\mathbf{V}^{1/2}\mathbf{W}\mathbf{V}^{1/2}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{1/2}\mathbf{W}\mathbf{V}^{1/2}\mathbf{z}.$$

[2] This approximation was suggested by Chambers (1973) in a related context for the analysis of more general nonlinear models.

This equation is identical to the corresponding exact noniterative solution for the standard linear model with $\mathbf{V}^{1/2}\mathbf{X}$ as the design matrix for the response variable $\mathbf{V}^{1/2}\mathbf{z}$. This is useful because the formulae displayed in Section 2 are then directly applicable to this one-step estimate. In particular, we find that

$$\hat{\boldsymbol{\beta}}^1(w) = \hat{\boldsymbol{\beta}} - \frac{(\mathbf{X}^T\mathbf{V}\mathbf{W})^{-1}\mathbf{x}_l s_l(1-w)}{\{1-(1-w)h_{ll}\}}. \tag{4}$$

Since the usual maximum likelihood fit corresponds to $w = 1$, decreasing w toward zero amounts to giving the lth point less and less weight in the fitting process. If the change in coefficients is negligible as w is decreased, then the lth observation exerts very little influence on the coefficients, and hence on the fit itself. On the other hand, if small changes in w induce large (relative) changes in $\hat{\boldsymbol{\beta}}$, then the lth point is influential, and we can proceed further to isolate which components of the fit are most unstable with respect to this point. The next section pursues this analysis.

6. Generalized regression diagnostics.

6.1. *Coefficient Sensitivity.* The above comments can conveniently be summarized by evaluating the derivative of $\hat{\boldsymbol{\beta}}^1(w)$ with respect to w at several reference points. In general, we have the result

$$\frac{\partial}{\partial w}\hat{\boldsymbol{\beta}}^1(w) = \dot{\hat{\boldsymbol{\beta}}}^1(w) = (\mathbf{X}^T\mathbf{V}\mathbf{W})^{-1}\mathbf{x}_l s_l/\{1-(1-w)h_{ll}\}^2.$$

As with linear regression, evaluation of $\dot{\hat{\boldsymbol{\beta}}}^1(w)$ at $w = 0, \bar{w}, 1$ provides differential information concerning the effect of the lth observation on the fit and can be interpreted as influence functions. For logistic regression, more of a case can be made for using $\dot{\hat{\boldsymbol{\beta}}}^1(\bar{w}) = \Delta_l\hat{\boldsymbol{\beta}}^1$ rather than the more liberal value $\dot{\hat{\boldsymbol{\beta}}}^1(0)$ since our one-step approximation may not be valid this far from the maximum. Computationally, using (4), the suggested diagnostic for individual coefficient sensitivity is given by

$$\Delta_l\hat{\boldsymbol{\beta}}^1 = (\mathbf{X}^T\mathbf{V}\mathbf{X})^{-1}\mathbf{x}_l s_l/(1-h_{ll}).$$

Plots of $\Delta_l\hat{\beta}_j^1/\text{s.e.}(\hat{\beta}_j)$ vs l are useful for the detection of observations that are causing instability in selected coefficients. For the example on vaso-constriction in the skin, these plots are displayed in Figure 6. The symbol (*) refers to the standardized difference in coefficients based on the fully iterated estimate $\hat{\boldsymbol{\beta}}(0)$, whereas the symbol ($o$) refers to the approximation to this quantity using $\hat{\boldsymbol{\beta}}^1(0)$. Excellent agreement is obtained in all but two cases, observations #4 and #18. Although the one-step estimates under-estimate the exact coefficient differences in these extreme cases, they clearly display the large effects that these points have on the estimates.

In the above example, it was clear that two observations had a large effect on all the estimated parameters. However, changes in particular coefficients can be offset by changes in other coefficients in such a way that the fitted values change very little, i.e.,

$$\frac{\partial}{\partial w}\hat{y}(w) = \frac{\partial}{\partial w}\dot{a}(\mathbf{X}\hat{\boldsymbol{\beta}}(w)) \doteq 0.$$

In other cases where a large number of explanatory variables are being fit, looking at index plots of each coefficient becomes unwieldly when it comes time to determine whether the observation in question has undue influence on the fit.

We approach this phase of the analysis by adapting an overall discrepancy measure due to Cook (1977). The motivation is as follows. The equation

$$-2\{l(\mathbf{X}\boldsymbol{\beta};\mathbf{y}) - l(\mathbf{X}\hat{\boldsymbol{\beta}};\mathbf{y})\} = c, \tag{5}$$

describes the boundary of an asymptotic confidence region for the parameter $\boldsymbol{\beta}$. Evaluation

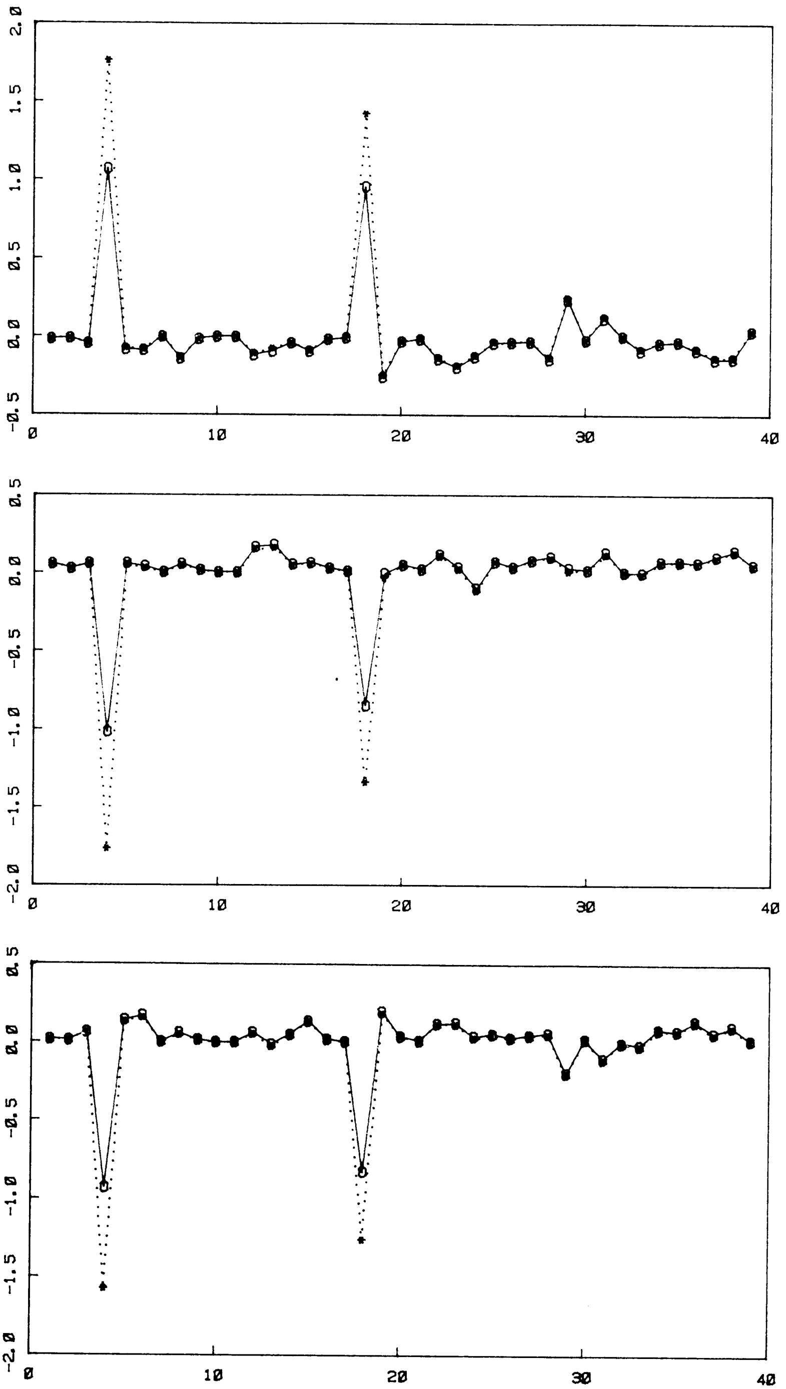

FIG. 6. *Index plots of the standardized change in coefficients for Finney's data: standardized change vs index number for $\hat{\beta}_1$ (top); $\hat{\beta}_2$ (middle); $\hat{\beta}_3$ (bottom). The symbols* (*) *and* (○) *respectively refer to the fully iterated and one-step standardized change.*

of this equation at $\beta = \hat{\beta}(0)$ leads to c_l, which gives a scalar measure of the influence of the lth point on the estimate $\hat{\beta}$. In particular, comparison of c_l to the percentage points of $\chi^2(m)$ gives a rough guide as to which contour of the confidence region the m.l.e. is displaced due to deleting the lth observation. Clearly one could extend this approach to infinitesimal perturbations rather than deletions.

In order to routinely compute c_l, an approximation is required. A second-order Taylor-series expansion of (5) leads to an approximate (ellipsoidal) confidence region, whose boundary is given by

$$(\beta - \hat{\beta})^T \mathbf{X}^T \mathbf{V} \mathbf{X} (\beta - \hat{\beta}) = c.$$

This corresponds to the confidence region determined by the limiting normal distribution of $\hat{\beta}$. Evaluating this equation at the one-step approximation to $\hat{\beta}(0)$ given by equation (4) leads to the confidence interval displacement diagnostic

$$c_l^1 = \frac{\chi_l^2 h_{ll}}{(1 - h_{ll})^2}. \tag{6}$$

Notice that c_l^1 is a function of the elementary building blocks of the previous section, all of which are available from the usual fit. This function is suggested for routine calculation following the usual fit. Again, an index plot, or a scatter plot versus $\hat{\theta} = \mathbf{x}\hat{\beta}$, should provide useful displays of these quantities.

Figure 7 displays the index plot for the skin vaso-constriction example. It is now apparent that the two observations in question have an undue influence on the fit itself, not just on the individual coefficients. The agreement between the exact and one-step versions of (6) is good, although we still tend to under-estimate the large effects.

Results similar to the above are available by considering the asymptotic confidence region for β defined by the inequality:

$$-2\{l(\mathbf{X}\beta; \mathbf{y}) - l(\mathbf{X}\hat{\beta}(0); \mathbf{y})\} \le \bar{c}. \tag{7}$$

Evaluation at $\beta = \hat{\beta}$ corresponds to the contour that $\hat{\beta}(0)$ is displaced by inclusion of the lth observation. Calculations similar to those corresponding to c_l^1 yield the confidence interval displacement diagnostic, which is given by

$$\bar{c}_l^1 = \chi_l^2 h_{ll}/(1 - h_{ll}).$$

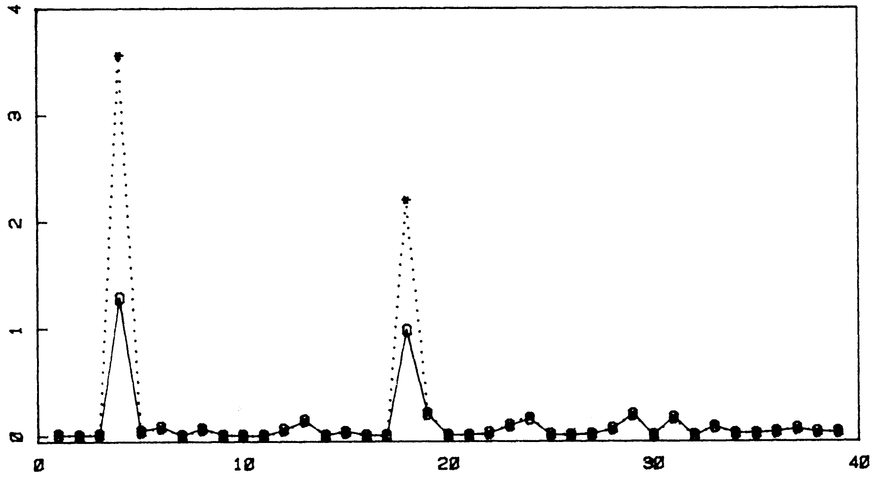

FIG. 7. *Index plot of confidence interval displacement for Finney's data: c^1 vs index number. The symbols (∗) and (O) respectively refer to the fully iterated and one-step confidence interval displacement.*

In comparison with c_l^1, this measure will necessarily be smaller in value. Essentially, $\bar{c}_l$ measures the overall change in fitted logits due to deleting the lth observation for all points excluding the one deleted. Conversely, c_l includes the deleted point. Although c_l will usually be the preferred diagnostic to measure overall coefficient changes, in the examples examined to date, the one-step approximations were more accurate for $\bar{c}_l^1$ than c_l^1.

The skin vaso-constriction data is a prime example of this behavior as evidenced by Figure 8. The index plot essentially conveys the same information as Figure 7, but the one-step approximation is much better in the present case.

6.2. *Goodness-of-Fit Sensitivity.* Another important aspect of the fit for which diagnostic information is valuable is the goodness-of-fit of the model itself. This can easily be studied by calculating the effect of changes in w on the statistics D and χ^2. Individual points which greatly influence the model are likely to induce large changes in the quality of the fit as judged by these statistics. We do not advocate using diagnostics on these statistics to create an artificial utopian model (e.g., by deleting all points which do not conform with specified standards), but rather to objectively identify points that do have substantial impact on the fit. Changes in these goodness-of-fit statistics can take two forms:

(1) if the lth point is not well fit by the model, the changes in D and χ^2 (caused by small changes in w) are usually (but as we shall see, not necessarily) isolated in the single components d_l and χ_l, and

(2) if the lth point is in an extreme or sparse region of the design space, the changes in D and χ^2 will be the result of all the individual components changing.

A large value of the change in these statistics will not indicate which of (1) or (2) is actually the case, but information from the other diagnostic tools (especially h_{ij}) should clarify the situation.

We proceed by first determining the effect on the deviance attributable to arbitrary values of w. Using the observation count device introduced in modifying the loglikelihood function, the corresponding deviance is

$$D_w(\mathbf{X}\hat{\beta}(w); \mathbf{y}) = 2 \textstyle\sum_{i=1}^{N} w_i\{l(\hat{\theta}_i; y_i) - l(\mathbf{x}_i\hat{\beta}(w); y_i)\}.$$

This is the deviance that would have resulted if we had actually observed data correspond-

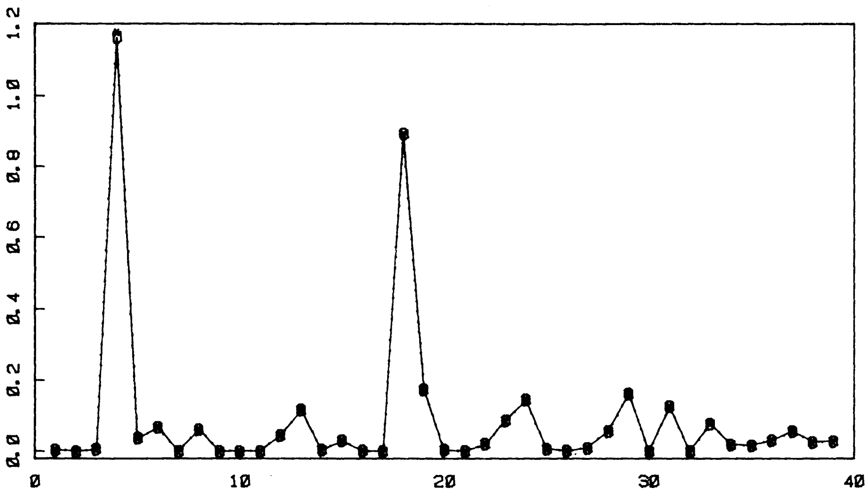

FIG. 8. *Index plot of confidence interval displacement for Finney's data: $\bar{c}^1$ vs index number. The symbols (*) and (O) respectively refer to the fully iterated and one-step confidence interval displacement.*

ing to the observation count given by w. Evaluation at the one-step estimate $\boldsymbol{\beta}^1(w)$ approximates this quantity. In particular, a second-order expansion of $D_w(\mathbf{X}\hat{\boldsymbol{\beta}}^1(w);\, \mathbf{y})$ about $\hat{\boldsymbol{\beta}}$ leads to

$$(8) \qquad D_w(\mathbf{X}\hat{\boldsymbol{\beta}}^1(w);\, \mathbf{y}) \doteq D(\mathbf{X}\hat{\boldsymbol{\beta}};\, \mathbf{y}) - [(1-w)d_l^2 + \chi_l^2(1-w)^2 h_{ll}/\{1-(1-w)h_{ll}\}].$$

This increasing function of w attains a minimum at $w = 0$ of $D(\mathbf{X}\hat{\boldsymbol{\beta}}^1;\, \mathbf{y}) - (d_l^2 + \bar{c}_l^1)$, and a maximum at $w = 1$ of $D(\mathbf{X}\hat{\boldsymbol{\beta}};\, \mathbf{y})$. That is, downweighting points *reduces* the deviance.

The rate of change of the deviance due to infinitesimal changes in w at the lth observation is easily obtained by differentiating (8). We again strike a compromise by evaluating at the one-step difference

$$\Delta_l D = D_1(\mathbf{X}\hat{\boldsymbol{\beta}};\, \mathbf{y}) - D_0(\mathbf{X}\hat{\boldsymbol{\beta}}^1(0);\, \mathbf{y}) \doteq d_l^2 + \frac{\chi_l^2 h_{ll}}{1 - h_{ll}}.$$

This gives the change (decrease) in deviance attributable to deleting the lth observation. Notice that all the basic building blocks play a role in this quantity, and all of them are available from the usual maximum likelihood fit. For display purposes an index plot of $\Delta_l D$, or the scatter plot of d_l^2 vs $\bar{c}_l^1$ are suggested. Alternatively, since

(a) $w = 0$ corresponds to deleting the lth observation,

(b) deleting the lth observation corresponds to augmenting $\mathbf{X}$ by a dummy variable $\mathbf{E}_l$, and

(c) the deviance reduction due to adding a variable is asymptotically $\chi^2(1)$, the individual values $\Delta_l D$ are asymptotically $\chi^2(1)$, and suitable for probability plotting. In particular, replacing χ_l by d_l in the definition of $\Delta_l D$ leads to $\Delta_l D \approx d_l^2/m_{ll}$ and the approximately normal studentized residual $d_l/\sqrt{m_{ll}}$. Accordingly we suggest using the ordered values of $d_l/\sqrt{m_{ll}}$ rather than of χ_l, as the ordinates for normal probability plots.

Determination of the effect of individual observations on χ^2 is more difficult. The underlying reason is that χ^2 is not as intimately related to the fitting procedure as D is. That is, D plays the same role for the logistic regression model as the RSS does for the normal linear regression model: the maximum likelihood (least-squares) estimate minimizes D (RSS). As observations are excluded from the fit, D (RSS) must decrease; χ^2 does not have this property, though only in extreme cases will χ^2 actually increase.

On the other hand, χ^2 is similar to RSS because both are sums of squares of deviations of observed from fitted values (in the unit normal model $D = \chi^2 =$ RSS). In fact, a one-step approximation to the change in χ^2 due to deleting the lth observation is given by

$$\Delta_l \chi^2 = \chi_l^2 - \chi_0^2 \doteq \frac{\chi_l^2}{1 - h_{ll}}.$$

which is the logistic regression analog of Δ_lRSS (see Section 2).

For the skin vaso-constriction example, index plots of $\Delta_l \chi^2$ and $\Delta_l D$ are displayed in Figure 9. The plots show that the 4th and 18th observations contribute heavily to the disagreement between the data and the fitted model. The one-step approximation to $\Delta_l D$ is excellent. The one-step approximation to $\Delta_l \chi^2$ is not very precise, though it clearly indicates the ill-fit points. This example is an extreme case where χ^2 actually increases as certain observations are deleted. In an effort to determine when and how this can occur, we now turn our attention to the effect of perturbations on the individual fitted values.

6.3. *Neighboring Effects.* We now derive a diagnostic measure which indicates the extent to which the lth point affects the fit at the remaining $N - 1$ points. A goal of this analysis is to determine how observations interact between themselves and the fit. This detailed analysis will usually be performed on a subset of $\lambda(\leq .05N)$ points that stood out (either clearly or marginally) in the previous diagnostic plots.

The naive approach considers the standardized change in the fit at the jth observation

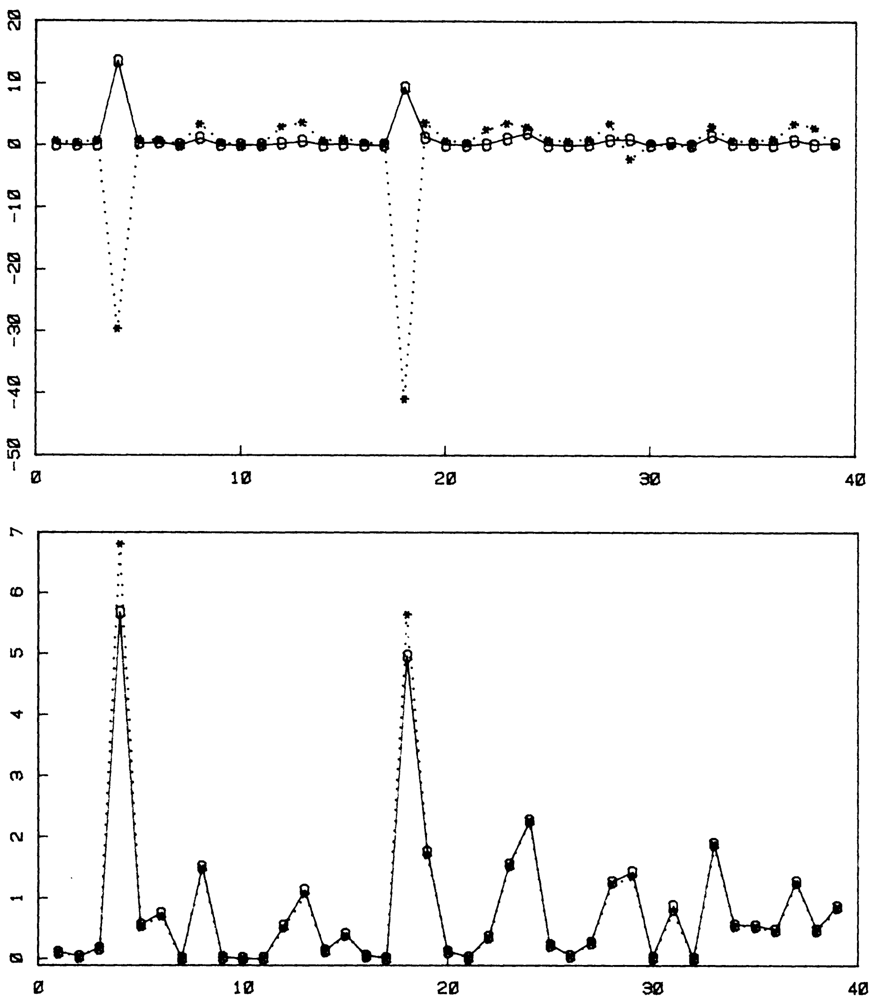

FIG. 9. *Index plots of change-in-fit diagnostics for Finney's data: change-in-fit diagnostic vs index number. The symbols (*) and (O) respectively refer to the fully iterated and one-step change-in-fit diagnostics.*

due to infinitesimal changes at the lth observation, i.e.,

$$\frac{\partial}{\partial w}\left\{\frac{\hat{y}_j(w)}{\text{s.e.}(\hat{y}_j)}\right\} \doteq \frac{h_{lj}\chi_l(1-w)}{\sqrt{h_{jj}}\{1-(1-w)h_{ll}\}}.$$

Note the presence of an off-diagonal element of **H**, such that, holding other things constant, large values of $| h_{lj} |$ imply large changes in fit. Further, since $| h_{lj} | \leq \sqrt{h_{ll}}\sqrt{h_{jj}}$, the absolute change in fit at j is bounded by that at l. At $w = 0$ this bound is$\chi_l\sqrt{h_{ll}}/(1 - h_{ll})$. In any case, whether we use the actual change or its absolute bound, this diagnostic can indicate the magnitude of change only, and not the direction. Since a large change in $\hat{y}_j$ can be the result of the fit moving toward y_j or away from it, this diagnostic does not give us an adequate view of the interaction between points. [See Cook (Section 4, 1977) for a related discussion for the linear regression model.]

It is also clear that examination of the h_{lj} alone will not produce much insight. This is so because the effect (either beneficial or detrimental) that one observation has on another depends not only on h_{lj}, but also on the position of the points in the residual space (i.e., the space orthogonal to $\mathbf{H}$). Since χ lies in the orthogonal complement of the column space of $\mathbf{H}$, we can expect the desired diagnostic to be a function of χ_l and χ_j, in addition to h_{lj}.

Consider the jth deviance contribution, $d_j^2(w)$, computed with $(w = 1)$ and without $(w = 0)$ the lth observation included in the fit:

$$\Delta_l d_j^2 = d_j^2(0) - d_j^2(1).$$

Note that

$\Delta_l d_j^2 > 0$ implies that the fit at the jth observation gets worse;
$\Delta_l d_j^2 = 0$ implies that the fit at the jth observation remains the same;
$\Delta_l d_j^2 < 0$ implies that the fit at the jth observation gets better.

Computationally, a one-step approximation to $\Delta_l d_j^1$ is given by

$$\Delta_l d_j^2 = \frac{2\chi_j h_{lj} \chi_l}{1 - h_{ll}} + \frac{\chi_l^2 h_{lj}^2}{(1 - h_{ll})^2},$$

a function of χ_j, h_{lj}, and χ_l as desired. The following three observations are important:

(1) the effect of l on j as measured by $\Delta_l d_j^2$ is not the same as the effect of j on l as measured by $\Delta_j d_l^2$. This is intuitively and geometrically reasonable;

(2) the fit at l cannot improve by deleting l from the fit as $\Delta_l d_l^2 \geq 0$; and

(3) the overall fit to the remaining $N - 1$ points *must* improve, as $\chi = \mathbf{M}\chi$ implies that $0 = \mathbf{H}\chi$, and $\Delta_l d_l^2 \geq 0$, leading to

$$\textstyle\sum_{j \neq l} \Delta_l d_j^2 = - \frac{\chi_l^2 h_{ll}}{1 - h_{ll}} \leq 0.$$

Any points with $\Delta_l d_j^2$ much greater than zero must be dealt with carefully because removing only the lth observation can mean disaster at these points. The most useful display for this purpose is the index plot of $\Delta_l d_j^2$ vs j for a subset of points indexed by $L = \{l_1, \ldots, l_\lambda\}$. Figure 10 displays this plot for the vaso-constriction data, and we immediately see the close connection between the 4th and 18th points. This indicates that one can improve the fit of the model only by dealing with these points together, and not separately. In comparison, the effect of the 31st observation on the fit at the remaining points is negligible.

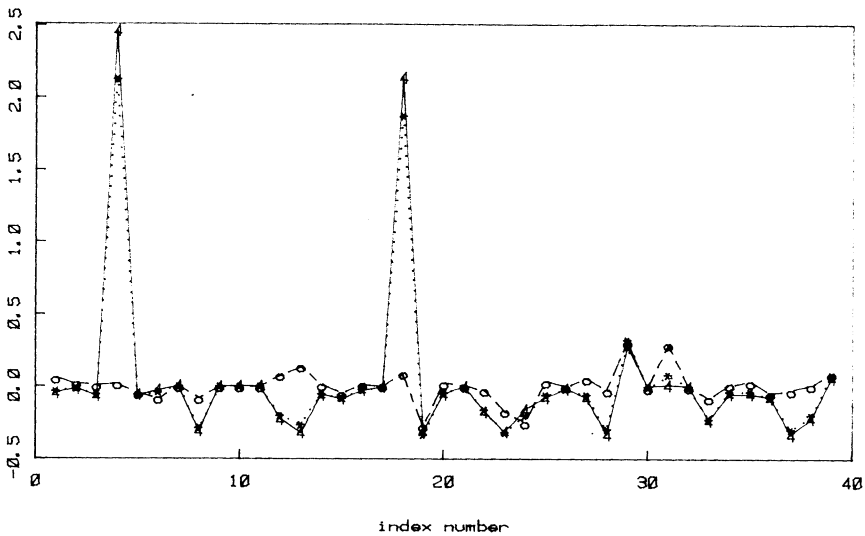

FIG. 10. *Index plot of neighboring effects for Finney's data: effect of observation l on j vs j, for observations l = 4(4), 18(*) and 31(0).*

7. Concluding remarks. As stated in Section 1, the methods employed here are easily extended to the class of generalized linear models; the notation was chosen to facilitate this extension. In general, "cheap" regression diagnostics can be derived for any model, provided that:

(1) one is given the fit of the model,

(2) one has access to the components of that fit, and

(3) one thoroughly understands the nature of the fitting process.

The methods of this paper are also easily extended to "subset" diagnostic measures, though we must recognize that here, we get nothing for free. In particular, if we let

$$w_i = \begin{cases} w & \text{for } \ i \in L \\ 1 & \text{otherwise} \end{cases}$$

where $L = \{l_1, l_2, \ldots, l_\lambda\}$ is an index set of observations, the methods follow directly. The main result is

$$\hat{\boldsymbol{\beta}}^1(w) = \hat{\boldsymbol{\beta}} - (\mathbf{X}^T\mathbf{V}\mathbf{X})^{-1}\mathbf{X}_L^T\mathbf{V}_L^{1/2}\{\mathbf{I} - (1-w)\mathbf{H}_L\}^{-1}\mathbf{V}_L^{-1/2}\mathbf{s}_L(1-w),$$

where any quantity with subscript L denotes the obvious components of that quantity indexed by the elements of L. Although we have the $\lambda \times \lambda$ matrix $\mathbf{H}_L$ as a by-product of the usual fit, it is necessary to compute $\{\mathbf{I} - (1-w)\mathbf{H}_L\}^{-1}$ to get diagnostic information on the λ-set in question. It is immediately apparent that much more work is required for subset diagnostics. The quality of the one-step estimate $\hat{\boldsymbol{\beta}}^1(w)$ relative to $\hat{\boldsymbol{\beta}}(w)$ is not so apparent. For details of the implementation of a subset analysis, see Pregibon (1979) or Belsley, Kuh, and Welsch (1980).

The formulae presented in this paper are based explicitly on the IRLS formulation of Section 3.2. These expressions also serve as computational formulae, though recent trends in numerical analysis suggest avoiding the computation of the sums-of-squares and products matrix (and its inverse) by using some accurate matrix decomposition method. If we use the QR decomposition, $\mathbf{V}^{1/2}\mathbf{X} = \mathbf{Q}\mathbf{R}$ with $\mathbf{R}$ $(m \times m)$ upper triangular and $\mathbf{Q}(N \times m)$ orthogonal $(\mathbf{Q}^T\mathbf{Q} = \mathbf{I})$, we have the basic result:

$$\hat{\boldsymbol{\beta}}^1(w) = \hat{\boldsymbol{\beta}} - \mathbf{R}^{-1}\mathbf{Q}_L^T\{\mathbf{I} - (1-w)\mathbf{Q}_L\mathbf{Q}_L^T\}^{-1}\chi_L(1-w)$$

which holds for singletons $L = \{l\}$ or subsets $L = \{l_1, l_2, \ldots, l_\lambda\}$. In the latter case, a carefully programmed algorithm would also avoid the direct inversion of $\mathbf{I} - (1-w)\mathbf{Q}_L\mathbf{Q}_L^T$. All diagnostics derived from this basic expression can be written in terms of $\mathbf{R}^{-1}$ and $\mathbf{Q}$.

The presence of the components of χ^2 in nearly every diagnostic is due to the nature of the one-step approximation, and not because we believe these residuals to be best. In particular, for the purposes of outlier detection and residual analysis, these quantities are unstable for either near-zero or near-unity fitted probabilities. We prefer the signed-deviance components for these purposes although other "denominatorless" χ^2-type residuals are often used (see for example, Mosteller and Tukey, 1977).

A very careful reader may have noticed a discrepancy between the data used here and those given in the Finney paper. The problem is that he reports $\text{RATE}_{32} = 0.03$ but $\log_{10}$ $(10 \times \text{RATE}_{32}) = 0.48$ $(=\log_{10}(10 \times 0.30))$. Judging from his plot of the data, RATE_{32} should be 0.30 rather than 0.03, but the latter value is the one that we used. Will this appreciably change the results of our analysis? *Hint*: Did observation 32 attract our attention in the diagnostic plots?

Acknowledgments. The author gratefully acknowledges the detailed comments that David Hoaglin provided on a preliminary draft. Roy Welsch pointed out a small error in a formula of Section 6 which would have otherwise gone undetected.

REFERENCES

[1] Andrews, D. F. and Pregibon, D. (1978). Finding the outliers that matter. *J. Royal Statist. Soc B* **40** 85–94.

[2] Belsley, D. A., Kuh, E., and Welsch, R. E. (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity.* Wiley, New York.

[3] Chambers, J. M. (1973). Fitting non-linear models: Numerical techniques. *Biometrika* **60** 1-13.

[4] Cook, R. D. (1977). Detection of influential observations in linear regression. *Technometrics* **19** 15-18.

[5] Cook, R. D. (1979). Influential observations in linear regression. *J. Amer. Statist. Assoc.* **74** 169-174.

[6] Cook, R. D. and Weisberg, S. (1979). Finding influential cases in linear regression: A review. Tech. Report 338, Univ. of Minnesota.

[7] Finney, D. J. (1947). The estimation from individual records of the relationship between dose and quantal response. *Biometrika* **34** 320-334.

[8] Hoaglin, D. C., and Welsch, R. E. (1978). The hat matrix in regression and ANOVA. *Amer. Statistician* **32** 17-22.

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

[10] Nelder, J. A., and Wedderburn, R. W. M. (1972). Generalized linear models. *J. Royal Statist. Soc. A* **135** 370-384.

[11] Pregibon, D. (1979). *Data Analytic Methods for Generalized Linear Models.* Unpublished Ph.D. thesis: Univ. of Toronto.

[12] Pregibon, D. (1980). Goodness of link tests for generalized linear models. *Appl. Statist.* **29** 15-24.

[13] Welsch, R. E., and Kuh, E. (1977). Linear regression diagnostics. Working Paper 173, Nat. Bur. Econ. Res. Inc.

Department of Biostatistics
School of Public Health and Community Medicine
University of Washington
Seattle, WA 98195