The Annals of Statistics
1984, Vol. 12, No. 3, 816–826

A GEOMETRIC APPROACH TO NONLINEAR REGRESSION DIAGNOSTICS WITH APPLICATION TO MATCHED CASE-CONTROL STUDIES[1]

BY SURESH H. MOOLGAVKAR, EDWARD D. LUSTBADER, AND DAVID J. VENZON

The Fox Chase Cancer Center, Philadelphia

A geometric approach is described for the detection of influential points in nonlinear regression. This work extends and gives geometric interpretation to recent results in logistic regression diagnostics. An application of this approach to matched case-control studies is discussed.

1. Introduction. A question of interest in the fitting of statistical models is how deletion of observations affects parameter estimates. In ordinary least squares with design matrix X and parameters β, an exact solution to the change in estimates is known, and the projection matrix $X(X^tX)^{-1}X^t$, also called the hat matrix, plays an important role (Hoaglin and Welsch, 1978). More recently, Pregibon (1981) considered logistic regression and derived an approximation to the change in the maximum likelihood estimate $\hat{\beta}$ on deletion of observations. Pregibon's procedure is based on one iteration of the Newton-Raphson algorithm, and by analogy with ordinary least squares, yields a "hat" matrix that plays the role of $X(X^tX)^{-1}X^t$ in linear regression. This procedure is easily extended to the exponential family when the natural parameter $\theta = X\beta$, i.e., when θ is a linear function of β.

This note presents a geometric construction of a hat matrix that is applicable to a wide range of problems in nonlinear regression. The matrix that we propose includes the usual hat matrix in ordinary least squares and Pregibon's hat matrix for the exponential family as special cases, and can be used to compute an approximation to the change in $\hat{\beta}$ on deletion of observations. However, when the model being fit is intrinisically nonlinear, our approach, in general, yields different approximations from the one-step Newton-Raphson approach, and it turns out that, for the exponential family, our diagnostics are identical to those based on one-step of the Fisher scoring algorithm, and are easier to implement. We also propose an appropriate definition of residuals for multivariate random variables, and a generalization of the notion of leverage.

The use of intrinsically nonlinear models is increasing in epidemiology, especially in the analysis of case-control studies by various extensions of the multiple logistic model. Our methods can be used to study the influence of entire risk sets

Received May 1983; revised March 1984.

[1] This work was supported by USPHS grants CA06927, CA22780 and CA30671 from the National Institutes of Health.

AMS 1980 *subject classifications*. Primary 62F99; secondary 62P10.

Key words and phrases. Case-control studies, diagnostics, hat matrix, influential observations, logistic regression.

and of individual controls. In order to use our geometric construction, we view the conditional likelihood of matched case-control studies as arising from a multinomial sampling scheme. This leads immediately to diagnostics for the elimination of entire risk sets. However, the computations are somewhat laborious. It turns out that a Poisson sampling scheme leads (up to a constant) to the same likelihood. With the Poisson model, the computation of the diagnostics is much simplified. Moreover, as is proved in the appendix, both models lead to identical diagnostics for elimination of entire risk sets, and the Poisson model (but not the multinomial) can be readily used to study the influence of individual controls. The equivalence of the diagnostics (for risk set deletion) based on the Poisson and multinomial schemes is particularly easy to prove in the geometric setting.

2. The exponential family and the hat matrix. Suppose that $Y_1, Y_2, \cdots, Y_n$ are independent vector valued random variables with Y_i being a member of the $m(i)$-parameter exponential family. Let $Y^t = (Y_1^t, Y_2^t, \cdots, Y_n^t)$ and let $y^t = (y_1^t, \cdots, y_n^t)$ be a realization of Y^t. The density of $Y_i^t = (Y_{i1}, \cdots, Y_{im(i)})$ is

$$g(y_i, \theta_i) = \exp\{\theta_i^t y_i - a(\theta_i) + b(y_i)\}$$

where $\theta_i^t = (\theta_{i1}, \cdots, \theta_{im(i)})$ is the vector of natural parameters. Let $\theta^t = (\theta_1^t, \theta_2^t, \cdots, \theta_n^t)$, and let β be a vector of parameters to be estimated. Since $\dot{a}(\theta) = f(\beta)$ is the expectation of Y, the normal equations are

$$(\partial\theta/\partial\beta)^t\{y - f(\beta)\} = 0. \tag{1}$$

When $\theta = X\beta$, the Newton-Raphson algorithm yields the following iterative scheme:

$$\hat{\beta}^{j+1} = \hat{\beta}^j + (X^t\hat{V}^jX)^{-1}X^t\hat{r}^j, \tag{2}$$

where $\hat{\beta}^j$ is the estimate at the jth iteration, $\hat{V}^j = \ddot{a}\{\theta(\hat{\beta}^j)\}$ is the (block diagonal) covariance matrix of Y at the jth iteration and $\hat{r}^j = \{y - f(\hat{\beta}^j)\}$. Let $\eta = X\hat{\beta} + \hat{V}^{-1}\hat{r}$, where $\hat{\beta}$ is the maximum likelihood estimate of β and $\hat{V}$ and $\hat{r}$ are evaluated at $\hat{\beta}$. Then, at convergence, expression (2) can be written as

$$\hat{\beta} = (X^t\hat{V}X)^{-1}X^t\hat{V}\eta. \tag{3}$$

Let $\hat{\beta}_{(k)}$ be the maximum likelihood estimate of β on deletion of the kth random variable. By analogy with a linear regression having design matrix $\hat{V}^{1/2}X$ and observation vector $\hat{V}^{1/2}\eta$, Pregibon (1981) estimates $\delta_k = \hat{\beta} - \hat{\beta}_{(k)}$ using the hat matrix

$$H = \hat{V}^{1/2}X(X^t\hat{V}X)^{-1}X^t\hat{V}^{1/2}. \tag{4}$$

This estimate of δ_k is identical to the one obtained by starting at $\hat{\beta}$ and taking one step towards $\hat{\beta}_{(k)}$ using the Newton-Raphson algorithm.

3. A geometric construction of the hat matrix. There are two important spaces associated with multivariate exponential families, the space of natural

parameters and the space of expectations. The geometry of these spaces, and, in particular, the geometry of maximum likelihood estimation has recently been the focus of some interest (Efron, 1975, 1978; Amari, 1982). The Fisher information matrix $-E(\partial^2\ell/\partial\theta^2) = V$ is a naturally defined inner product on the space of natural parameters, whereas its inverse is the natural inner product on the dual space, i.e., on the space of expectations. Maximum likelihood estimation may be viewed geometrically in the space of expectations as follows.

As before, $f(\beta) = \{f_1^t(\beta), f_2^t(\beta), \cdots, f_n^t(\beta)\}^t$ is the vector of expectations of $Y = (Y_1^t, \cdots, Y_n^t)^t$ as a function of β. The quadratic form $\{y - f(\beta)\}^t V^{-1} \cdot \{y - f(\beta)\}$ is minimized at $\hat{\beta}$ when the Y_i are in the exponential family. Geometrically, the weighted distance between y and the locus of expectations $f(\beta)$ is minimized at $f(\hat{\beta})$. Thus, the vector $\{y - f(\hat{\beta})\}$ is perpendicular to the tangent plane of the locus of $f(\beta)$ at $f(\hat{\beta})$ in the inner product defined by $\hat{V}^{-1}$ (note that $-E(\partial^2\ell/\partial\theta^2)_{\theta=\theta(\hat{\beta})} = \hat{V}$). This is true even if $\theta = h(\beta)$ is not a linear function of β. Let J_β denote the Jacobian matrix of f at β. The columns of $J_{\hat{\beta}}$ span the vector space parallel to the tangent plane at $f(\hat{\beta})$, and the orthogonal (in the inner product defined by $\hat{V}^{-1}$) projection onto this vector space is defined by the matrix

$$\tilde{H} = \hat{V}^{-1/2} J_{\hat{\beta}} (J_{\hat{\beta}}^t \hat{V}^{-1} J_{\hat{\beta}})^{-1} J_{\hat{\beta}}^t \hat{V}^{-1/2}. \tag{5}$$

Note that the columns of $\hat{V}^{1/2}$ are orthonormal vectors for the inner product defined by $\hat{V}^{-1}$. The matrix $\tilde{H}$ is expressed in terms of the basis consisting of these orthonormal vectors. The geometric view of $\tilde{H}$ as a projection matrix suggests its use as a hat matrix for any generalized least squares problem. This approach is equivalent to using the linear least squares problem $y = J_{\hat{\beta}}\beta + r$, $E(r) = 0$ and $\text{var}(r) = \hat{V}$, to estimate δ_k in the nonlinear problem. The appropriate formula is given by (6) below.

Let $J_{\hat{\beta}k}$, $\tilde{H}_k$, and $\hat{V}_k$ be the Jacobian matrix, the hat matrix, and the block-diagonal covariance matrix, respectively, restricted to the kth random variable. Then,

$$\delta_k = \hat{\beta} - \hat{\beta}_{(k)} \approx (J_{\hat{\beta}}^t \hat{V}^{-1} J_{\hat{\beta}})^{-1} J_{\hat{\beta}k}^t \hat{V}_k^{-1/2} (I - \tilde{H}_k)^{-1} \hat{V}_k^{-1/2} \hat{r}_k, \tag{6}$$

where $\hat{r}_k = y_k - f_k(\hat{\beta})$. Note that $J_{\hat{\beta}}^t \hat{V}^{-1} J_{\hat{\beta}}$ is the expected (Fisher) information matrix at $\hat{\beta}$.

If $\theta = X\beta$, i.e., if θ is a linear function of β, then an easy computation shows that $J_{\hat{\beta}} = \hat{V}X$, and substitution of this into (5) yields (4). Thus, the approximation (6) is the same as the one-step Newton-Raphson approximation suggested by Pregibon. When $\theta = h(\beta)$ is not a linear function of β, then $J_{\hat{\beta}} = \hat{V}X_{\hat{\beta}}$ where X_β is the matrix $\partial\theta/\partial\beta$, and (5) yields (4) with X replaced by $X_{\hat{\beta}}$. In this case, the one-step approximation to δ_k based on the Newton-Raphson algorithm does not yield (6). Moreover, the one-step Newton-Raphson approximation is more difficult to implement than (6). This is easily seen by noting that derivatives of the left-hand side of the normal equations (1) will involve products of X_β and $f(\beta)$. However, in all cases, (6) can be interpreted as being a one-step estimate based on the Fisher scoring algorithm for maximum likelihood estimation.

The above discussion has been restricted to the exponential family. However,

in any weighted least squares problem, the hat matrix can still be constructed as in (5) with $\hat{V}^{-1}$ replaced by the appropriate positive definite weight matrix. In this case, (6) can be interpreted as being a one-step estimate based on the Gauss-Newton algorithm for nonlinear least squares estimation (see, e.g., Cook and Weisberg, 1982, pages 186–187).

4. The diagnostics. Expression (6) suggests three additional diagnostic measures to evaluate the impact of the kth observation.

The quadratic form $\hat{r}_k^t \hat{V}_k^{-1} \hat{r}_k$ will be denoted by $|\hat{r}_k|^2$ and provides a natural one-dimensional summary residual statistic. Since

$$\{y - f(\hat{\beta})\}^t \hat{V}^{-1} \{y - f(\hat{\beta})\} = \sum_{k=1}^n |\hat{r}_k|^2,$$

a plot of $|\hat{r}_k|^2$ can be used as a diagnostic like the familiar residual plots in linear regression with the obvious difference that $|\hat{r}_k|^2$ is always positive.

Another diagnostic of interest is an analogue of the Cook distance (Cook, 1977)

$$D_k = \{(\hat{\beta} - \hat{\beta}_{(k)})^t (J_{\hat{\beta}}^t \hat{V}^{-1} J_{\hat{\beta}})(\hat{\beta} - \hat{\beta}_{(k)})\}^{1/2}. \tag{7}$$

D_k is a standardized measure of how much deletions of observations affect the parameter values. We refer to D_k as the influence of Y_k.

Finally, let $M_k = I - \tilde{H}_k$. Then it is clear from the expression (6) that small values of $\det M_k$ imply large changes in parameter estimates. The quantity $1 - \det M_k$ will be called the leverage of Y_k. If Y_k is one-dimensional, this reduces to the usual notion of leverage.

5. The matched case-control study. Logistic regression is assuming increasing importance in the analysis of matched case-control studies. Usually, the odds ratio $\rho(\beta, x) = \exp(\beta^t x)$. However, other functional forms for the odds ratio are of interest, and Thomas (1981) has considered various general relative risk models. A question of interest is how do individual risk sets and individual controls within risk sets affect the fit of the model and the parameter estimates?

Suppose that there are n cases, and that for the ith case there are $m(i)$ controls. Let R_i denote the ith risk set, i.e., the ith case together with its controls. Whatever the form of $\rho(\beta, x)$, the appropriate conditional likelihood function is

$$L = \prod_{i=1}^n \rho(\beta, x_{i0}) / \sum_{q=0}^{m(i)} \rho(\beta, x_{iq}) \tag{8}$$

where x_{iq} is the vector of covariates for the qth individual in risk set i, and $q = 0$ corresponds to the case (Thomas, 1981).

The likelihood (8) can also arise by letting $Y_1, Y_2, \cdots, Y_n$ be independent random variables such that Y_i has a multinomial distribution with cell probabilities $(P_{i0}, \cdots, P_{im(i)})$ and realization $y_i = (1, 0, \cdots, 0)$, where

$$P_{is} = \rho(\beta, x_{is}) / \sum_{q=0}^{m(i)} \rho(\beta, x_{iq}). \tag{9}$$

It is easily seen that (8) is the product of the P_{i0} which is the likelihood function for this multinomial realization. The natural parameters of Y_i as a member of

the $m(i)$-parameter exponential family are

$$\theta_{is} = \log(P_{is}/P_{i0}) = \log \rho(\beta, x_{is}) - \log \rho(\beta, x_{i0}), \quad s = 1, \cdots, m(i).$$

Thus, θ is a linear function of β if, and only if, $\rho(\beta, x_{is}) = \exp(\beta^t x_{is})$.

As noted in Section 3, the Jacobian matrix is given by $V_\beta X_\beta$, where V_β is the block diagonal covariance marix of the Y_i and X_β is the matrix of first derivatives $\partial\theta/\partial\beta$. The hat matrix may then be constructed as in (5).

There are computational problems in working with this hat matrix since it necessitates taking the square root of a block diagonal covariance matrix. Moreover, while this hat marix can be used to study the effect of deleting entire risk sets, it is not practical to study the effect of individual controls within risk sets. These problems can be resolved by letting the P_{is} in (9) be the expectations of $m(i) + 1$ independent Poisson variables. For this model, the likelihood contribution from risk set R_i is

$$P_{i0}\exp(-P_{i0}) \prod_{s=1}^{m(i)} \exp(-P_{is}) = \exp(-1)P_{i0}.$$

The hat matrix for this model may also be constructed as in (5). The matrix $\hat{V}$ is now the diagonal matrix of variances of the Poisson random variables.

Estimates of δ_k on dropping risk sets, $|\hat{r}_k|^2$, the Cook distance (influence), and leverage may be computed as in Section 4. Although it seems plausible, it does not follow directly from the equivalence of the likelihoods that, for the deletion of an entire risk set, these diagnostics are identical for the Poisson and multinomial models. For our geometric approach, they are, as is discussed in the appendix.

An approximation to the changes in parameter estimates for omission of individual controls may also be computed with an expression like (6) based on the Poisson approach. However, if one control is deleted from risk set k, the estimated probabilities for the case and the remaining $m(k) - 1$ controls do not sum to 1. It is our experience that the approximations to the change in parameter estimates are better if the probabilities are adjusted to sum to 1. Using these adjusted probabilities, an appropriate diagnostic for the change in the estimate of β after deleting control s in risk set k is

$$\begin{aligned}\hat{\beta} - \hat{\beta}_{(ks)} &\approx (J_{\hat{\beta}}^t \hat{V}^{-1} J_{\hat{\beta}})^{-1} J_{\hat{\beta}ks}^t \hat{v}_{ks}^{-1/2}(1 - h_{ks})^{-1}\hat{v}_{ks}^{-1/2}\tilde{r}_{ks} \\ &= -(J_{\hat{\beta}}^t \hat{V}^{-1} J_{\hat{\beta}})^{-1} J_{\hat{\beta}ks}^t / \{(1 - \hat{P}_{ks})(1 - h_{ks})\}\end{aligned}$$

where $J_{\hat{\beta}ks}$, v_{ks}, and h_{ks} are the row of $J_{\hat{\beta}}$, the diagonal element of $\hat{V}$ and the diagonal element of $\tilde{H}$, respectively, corresponding to that control, and where $\tilde{r}_{ks} = \hat{r}_{ks}/(1 - \hat{P}_{ks}) = -\hat{P}_{ks}/(1 - \hat{P}_{ks})$.

6. An example. In this section, we discuss an application of the diagnostics to a matched case-control study of endometrial cancer in Los Angeles (Mack et al., 1976). The data are also presented in Appendix III of Breslow and Day (1980), and are available from us on request. There are 63 cases (63 risk sets) with 4 controls per case in that data set. Two covariates, one of them discrete (presence or absence of gall bladder disease), and the other continuous (length of estrogen use in months), were chosen for analysis. We eliminated 6 risk sets because these

had a missing covariate value for the case. In addition, we eliminated 1 control from each of 8 other risk sets because a covariate value was missing on that control. Thus, we analyzed a total of 277 observations arranged in 49 risk sets consisting of a case and 4 controls each, and 8 risk sets consisting of a case and 3 controls each. An additive formulation was used for the relative risk $\rho(\beta, x) = 1 + \beta_1 x_1 + \beta_2 x_2$. For this model, our diagnostics are different from those based on the Newton-Raphson one-step approximation.

In all that follows, we will be using the Poisson model for the matched case-control study. The hat matrix $\hat{H}$ is defined as in (5) and is a 277 × 277 matrix with rank 2 (since number of covariates = 2). For any risk set with $\hat{r}_k = (1 - \hat{P}_{k0}, -\hat{P}_{k1}, \cdots, -\hat{P}_{km})$, where $m = 3$ or 4 and $\hat{P}_{ks}$ are the fitted values (probabilities), each $\hat{H}_k$ is a 4 × 4 or 5 × 5 matrix.

Since the expression (6) for δ_k is an approximation, it is crucial to evaluate whether the approximation is adequate to draw attention to risk sets that lead to large changes in parameter estimates when dropped. Figure 1 shows the actual change in parameter estimates, and the approximate change, using both one-step

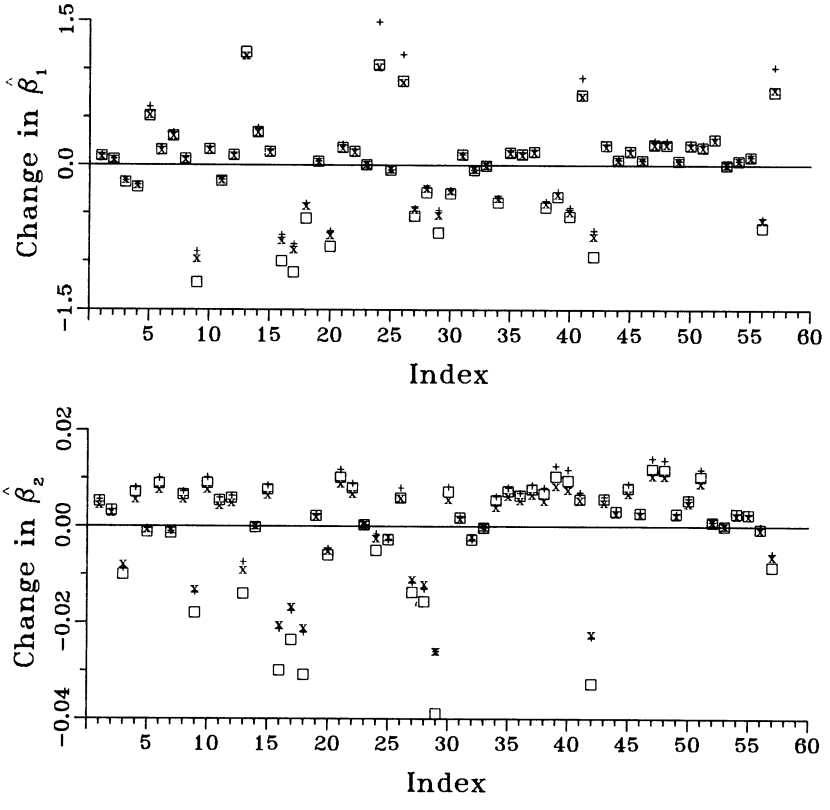

FIG. 1. *Change in parameter estimates on deleting entire risk sets. Index refers to the index of the risk set deleted. An open box denotes the actual change in parameter estimates, a "+" denotes the Newton-Raphson one-step approximation, and an "x" denotes the approximation based on our geometric approach. Maximum likelihood estimates of the parameters with all risk sets included are* $\hat{\beta}_1 = 4.26$, $\hat{\beta}_2 = 0.12$.

Newton-Raphson and our approach. In this example, expression (6) does somewhat better than the one-step Newton-Raphson approximation, and, contrary to Pregibon's (1981) experience, the latter occasionally overestimates the change in coefficients. We judge the approximation by (6) to be conservative, but adequate to draw attention to unusual risk sets. Expression (6) involves the inversion of $(I - \tilde{H}_k)$. However, matrix inversion may be avoided by noting that each $\tilde{H}_k$ is a submatrix of a large projection matrix of rank 2. This implies that $\tilde{H}_k$ is small and $(I - \tilde{H}_k)^{-1} = \sum_{j=0}^{\infty} (\tilde{H}_k)^j$, with $(\tilde{H}_k)^0 = I$. In our computations, we use the approximation $(I - \tilde{H}_k)^{-1} \approx I + \tilde{H}_k$, which does an excellent job.

An easy computation shows that $|\hat{r}_k|^2 = (1 - \hat{P}_{k0})/\hat{P}_{k0}$. Nominal levels for $|\hat{r}_k|^2$ that are based on distributional properties are difficult to compute. However, a simple computation shows that, conditional on the risk sets, the expectation of $|\hat{r}_k|^2$ is equal to the number of controls in the risk set. This is also the value of $|\hat{r}_k|^2$ under the null hypothesis that $\beta = 0$. Thus, risk sets with $|\hat{r}_k|^2$ larger than the number of controls are worse fit by the model under consideration than by the null model, and a plot of $|\hat{r}_k|^2$ should serve to draw attention to these risk sets.

Of course, in any analysis of matched case-control data, some of the risk sets are likely to have values of $|\hat{r}_k|^2$ larger than the nominal level we have proposed. To take an extreme example, in a matched pair analysis ($m(i) = 1$), using the usual logistic model and with one binary covariate ($d = 1$), let $a_{1,0}$ be the number of pairs with the covariate present for the case and absent for the control, and

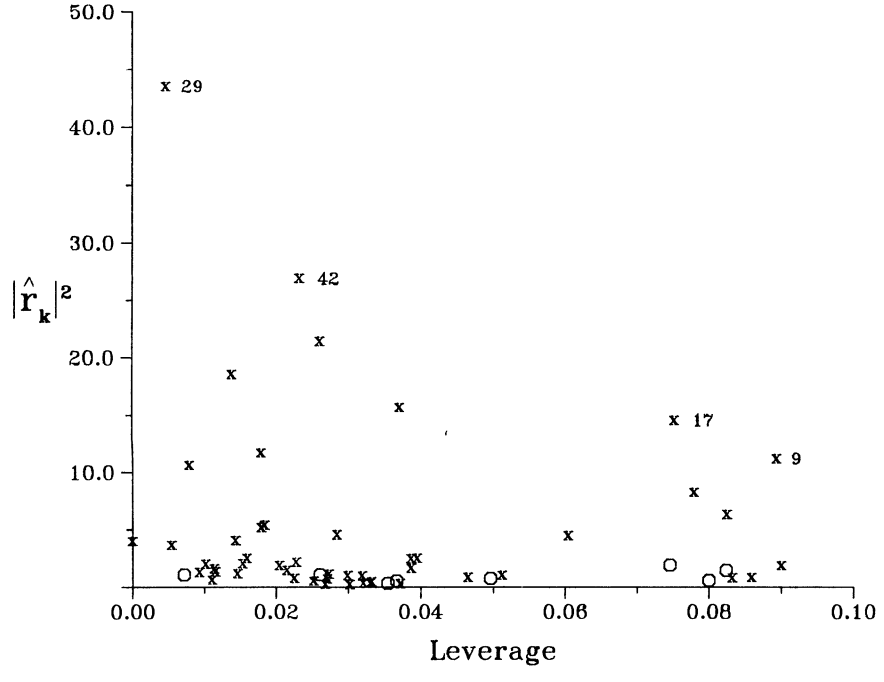

FIG. 2. *Plot of* $|\hat{r}_k|^2$ *versus leverage. Note that one risk set* (29) *has a very large value of* $|\hat{r}_k|^2$. *A few risk sets have both high leverage and high* $|\hat{r}_k|^2$. *An "x" denotes a risk set with* 4 *controls, an "o" denotes a risk set with* 3 *controls.*

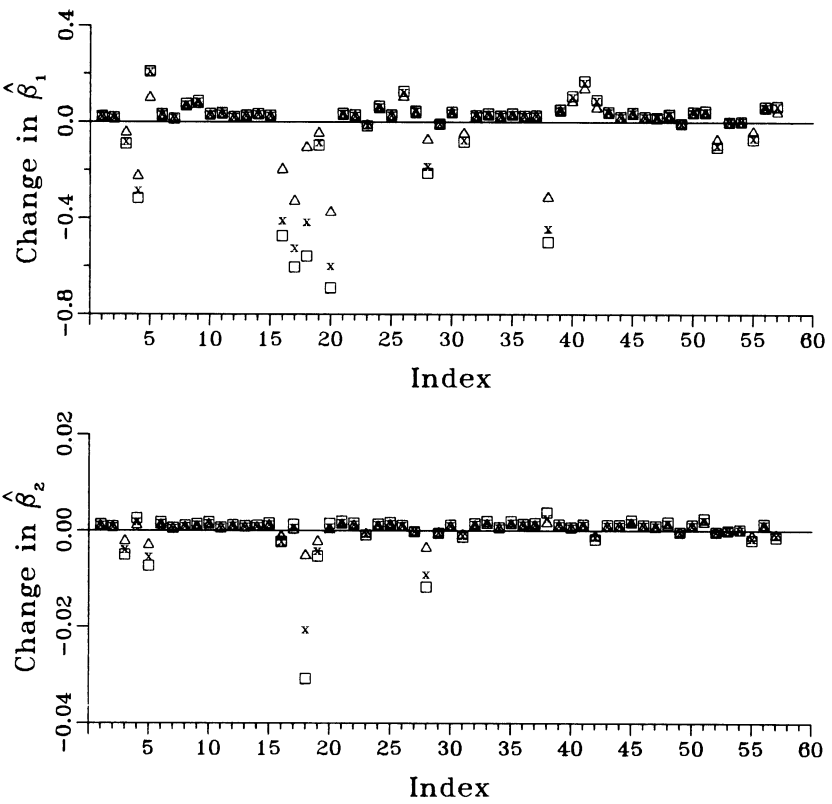

FIG. 3. *Change in parameter estimates on deleting individual controls. Index refers to the index of the risk set; the first control in each risk set has been dropped. An open box denotes the actual change in parameter estimates, an "x" denotes the approximate change with the correction factor* $(1 - \hat{P}_{k1})^{-1}$ *and an open triangle denotes the approximate change without the correction factor. Maximum likelihood estimates with all individuals included are* $\hat{\beta}_1 = 4.26$, $\hat{\beta}_2 = 0.12$.

$a_{0,1}$ be the number of pairs with the covariate absent for the case and present for the control. The estimate of the odds ratio is $a_{1,0}/a_{0,1}$, and thus determines the number of risk sets with $|\hat{r}_k|^2 > 1$, which is the nominal value in this case. The statistic $|\hat{r}_k|^2$ will be most useful when the number of covariates is large and there are both continuous and discrete covariates.

Large values of leverage for a risk set indicate an unusual combination of covariates in that risk set. Since rank $(\tilde{H}) = 2$, the average diagonal entry is $2/277 \approx 0.007$, and the average trace of a 5×5 matrix $\tilde{H}_k$ is 0.035. For each of these $\tilde{H}_k$, there are, at most, 2 nonzero eigenvalues and, therefore, the "average" size of such an eigenvalue is 0.018. Now each $M_k = I - \tilde{H}_k$ has 5 eigenvalues. Of these, three are equal to 1 (corresponding to the zero eigenvalues of $\tilde{H}_k$); the other two lie in the closed interval [0, 1]. On average, these two eigenvalues of M_k are $1 - 0.018 = 0.982$, and the average determinant is $(0.982)^2 \approx 0.964$; hence, values of $1 - \det M_k$ greater than 0.036 indicate risk sets with higher than average leverage.

Figure 2 is a plot of $|\hat{r}_k|^2$ versus leverage. Risk set 29 stands out as having

very high $|\hat{r}_k|^2$. A few risk sets (e.g., 9, 17) have both high leverage and high $|\hat{r}_k|^2$.

Figure 3 is a plot of the change in parameter estimates on dropping individual controls. The first control in each risk set was dropped for this figure. The correction factor $(1 - \hat{P}_{k1})^{-1}$ improves the approximation considerably.

Finally, although a convenient expression like (6) is unavailable, for the example discussed here, the one-step Newton-Raphson approximation was relatively easy to implement because $\rho(\beta, x)$ is linear in β. However, there is some interest in using more general forms of the relative risk function $\rho(\beta, x)$. For example, Thomas (1981) advocates using a mixture model in which

$$\rho(\alpha, \beta, x) = (1 + \beta^t x)^{1-\alpha}\{\exp(\beta^t x)\}^{\alpha}.$$

In such instances, expression (6) based on the geometric approach is considerably easier to implement than the one-step Newton-Raphson approximation.

7. Concluding remarks. We have presented a simple geometric approach to diagnostics for general nonlinear regression. Our approach is identical to a one-step approximation based on the Gauss-Newton algorithm. For the exponential family, this algorithm corresponds to the method of scoring for parameters which is essentially the Newton-Raphson algorithm with the expected information matrix taking the place of the observed one. A geometric view of one-step approximations based on these two algorithms (Newton-Raphson and scoring for parameters) is as follows. As noted earlier, the expected information matrix (with the kth observation deleted) defines an inner product on the space of parameters. Similarly, in a small enough neighborhood of $\hat{\beta}_{(k)}$, the observed information matrix (with the kth observation deleted) also defines an inner product. Thus, if $\hat{\beta}$ is close enough to $\hat{\beta}_{(k)}$, the observed information matrix defines an inner product in a neighborhood of $\hat{\beta}$. The gradients (Milnor, 1963, page 12) of the log likelihood with respect to these inner products can be defined at $\hat{\beta}$. The one-step approximations then consist of adding the appropriate gradient vectors to $\hat{\beta}$.

In the exponential family, when $\theta = X\beta$, the diagnostics presented here are identical to those suggested by Pregibon (1981). However, the approach described here is applicable to any generalized least squares problem. For example, the methods proposed in this paper are directly applicable to what Wedderburn (1974) calls maximum quasilikelihood estimation.

For matched case-control studies, our approach yields diagnostics for a general relative risk formulation. Moreover, the Poisson model provides diagnostics for deletion of individual controls.

We note here that the partial likelihood (Cox, 1975) that arises in the analysis of survival data via the proportional hazards model is formally identical to (8). Thus, the approach described here is applicable. However, because of the nesting of risk sets, the diagnostics for the deletion of entire risk sets are difficult to interpret. It is more desirable to compute the diagnostics for deleting individuals. Most individuals in a survival study appear in several risk sets. In order to implement the formula (6), it is necessary to keep track of the contribution made

by the individual in each risk set so that the matrices $J_{\hat{\beta}k}$, $\hat{V}_k$, $\hat{H}_k$, and the vector $\hat{r}_k$ can be computed. The procedure is illustrated in a recent manuscript (Lustbader and Moolgavkar, 1984). The risk sets are typically much larger than the risk sets for case-control studies, and the computational problems are correspondingly more difficult. In particular, (6) involves the inversion of a large matrix. However, as noted above, since $\hat{H}_k$ is a submatrix of a projection matrix, $(I - \hat{H}_k)^{-1}$ may be replaced by the first few terms of the series $\sum_{j=0}^{\infty} (\hat{H}_k)^j$. In our experience, $I + \hat{H}_k$ does an excellent job.

In this paper, we have restricted our attention to risk sets with only single failures. When there are several failures in a risk set, our method is easily generalized to yield diagnostics for deletion of entire risk sets. In the context of survival analysis, if the ties are broken by the method of Peto (1972) and Breslow (1972), then our approach will yield diagnostics for deleting individuals. However, if there are several failures per risk set, breaking of ties is not to be recommended. This situation could arise in a stratified case-control study, and there is no obvious generalization of our method to yield deletion diagnostics for single individuals. However, in this situation, the deletion of entire strata is likely to be of primary interest, and this can be handled by our methods as noted above.

Finally, yet another approach to diagnostics is advocated by Storer and Crowley (1984). They note that an approximation to δ_k may be obtained by fitting an augmented regression model. Prentice (personal communication) has also been exploring a similar approach. Pregibon (1984) has explored the Newton-Raphson one-step approximation in the context of matched case-control studies. Ultimately, the choice among various approaches will have to be made on the basis of convenience, computational ease, and the adequacy of the approximation.

APPENDIX

In this appendix, we sketch a proof of the proposition that the multinomial and Poisson approaches yield identical diagnostics (Section 4) and estimates of $\hat{\beta} - \hat{\beta}_{(k)}$ for the elimination of risk sets. Let $Y^t = (Y_0, \cdots, Y_m)$ be a multinomial random variable with parameters $(1, P_0, P_1, \cdots, P_m)$, $\sum_j P_j = 1$. Let $Z^t = (Z_0, \cdots, Z_m)$ be a vector of $m + 1$ independent Poisson random variables with expectation vector $(P_0, \cdots, P_m)$. Let $\beta \in R^d$ be a parameter, $f_j(\beta) = P_j$, $f(\beta) = (f_1(\beta), \cdots, f_m(\beta))^t$, and $g(\beta) = (f_0(\beta), f_1(\beta), \cdots, f_m(\beta))^t$. Let J_1 be the Jacobian matrix of f and J_2 the Jacobian matrix of g. Then, J_2 is a linear transformation from R^d into $U \subset R^{m+1}$ where U is the subspace defined by $\sum_{j=0}^{m} x_j = 0$. Let V_1 be the $m \times m$ covariance matrix of $(Y_1, Y_2, \cdots, Y_m)$ and let V_2 be the $(m + 1) \times (m + 1)$ diagonal matrix of variances of $(Z_0, Z_1, \cdots, Z_m)$. Then, V_1^{-1} and V_2^{-1} define inner products on R^m and R^{m+1}, respectively. Let $\{e_j\}_{j=1,\ldots,m}$ be the standard Euclidean basis for R^m, and let the basis for U consist of the set of vectors $\{u_j\}_{j=1,\ldots,m}$, with $u_j = (-1, 0, \cdots, 1, 0, \cdots, 0)$, -1 in the 0th position, 1 in the jth position. Then, the linear transformation $A: R^m \to U$, with $A(e_j) = u_j$, is an isometry. Let J_1^* and J_2^* denote, respectively, the adjoints for J_1 and J_2 with the standard inner product on R^d and the appropriate inner product (V_1^{-1} or V_2^{-1}) on the image space (R^m or R^{m+1}). (For a definition of adjoint, see Hoffman and Kunze, 1971, page 295.) The appropriate hat matrices (5) for the multinomial

and Poisson models are then $J_1(J_1^*J_1)^{-1}J_1^*$ and $J_2(J_2^*J_2)^{-1}J_2^*$, respectively. Under the isometry A, R^m and U are identical as inner product spaces, and the above two hat matrices can be identified with each other. Our proposition now follows easily.

Further details can be found in a recent technical report (Moolgavkar, Lustbader, and Venzon, 1982).

REFERENCES

AMARI, S. I. (1982). Differential geometry of curved exponential families—curvatures and information loss. *Ann. Statist.* **10** 357–385.

BRESLOW, N. E. (1972). Contribution to the discussion of the paper by D. R. Cox. *J. Roy. Statist. Soc., Ser. B* **34** 216–217.

BRESLOW, N. E. and DAY, N. E. (1980). *Statistical Methods in Cancer Research. Vol. 1—The Analysis of Case-Control Studies.* International Agency for Research on Cancer, Lyon.

COOK, R. D. (1977). Detection of influential observations in linear regression. *Technometrics* **19** 15–18.

COOK, R. D. and WEISBERG, S. (1982). *Residuals and Influence in Regression.* Chapman, New York.

COX, D. R. (1975). Partial likelihood. *Biometrika* **62** 269–276.

EFRON, B. (1975). Defining the curvature of a statistical problem (with applications to second order efficiency). *Ann. Statist.* **3** 1189–1242.

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

HOAGLIN, D. C. and WELSCH, R. G. (1978). The hat matrix in regression and ANOVA. *Amer. Statist.* **32** 17–22.

HOFFMAN, K. and KUNZE, R. (1971). *Linear Algebra.* Prentice-Hall, London.

LUSTBADER, E. D. and MOOLGAVKAR, S. H. (1984). A diagnostic statistic for the score test. Unpublished.

MACK, T. H., PIKE, M. C., HENDERSON, B. E., PFEFFER, R. I., GERKINS, V. R., ARTHUR, B. S. and BROWN, S. E. (1976). Estrogens and endometrial cancer in a retirement community. *New Engl. J. Med.* **294** 1262–1267.

MILNOR, J. (1963). Morse theory. *Ann. Math. Studies* **51**. Princeton University Press, Princeton, N.J.

MOOLGAVKAR, S. H., LUSTBADER, E. D. and VENZON, D. J. (1982). A geometric approach to non-linear regression diagnostics with application to matched case-control studies. University of Washington, Department of Biostatistics, Technical Report No. 53.

PETO, R. (1972). Contribution to the discussion of the paper by D. R. Cox. *J. Roy. Statist. Soc., Ser. B.* **34** 216–217.

PREGIBON, D. (1981). Logistic regression diagnostics. *Ann. Statist.* **9** 705–724.

PREGIBON, D. (1984). Data analytic methods for matched case-control studies. *Biometrics* (in press).

STORER, B. E. and CROWLEY, J. (1984). A diagnostic for Cox regression and general conditional likelihoods. *J. Amer. Statist. Assoc.* (in press).

THOMAS, D. C. (1981). General relative-risk models for survival time and matched case-control analysis. *Biometrics* **37** 673–686.

WEDDERBURN, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika* **61** 439–447.

EDWARD D. LUSTBADER
DAVID J. VENZON
THE FOX CHASE CANCER CENTER
7701 BURHOLME AVENUE
PHILADELPHIA, PENNSYLVANIA 19111

SURESH H. MOOLGAVAKAR
THE FRED HUTCHINSON CANCER CENTER
1124 COLUMBIA STREET
SEATTLE, WASHINGTON 98104