Austral. J. Statist., **32**(1), 1990, 11–20

INFLUENCE DIAGNOSTICS FOR THE NORMAL LINEAR MODEL WITH CENSORED DATA

L.A. WEISSFELD[1] AND H. SCHNEIDER[2]

University of Michigan and Louisiana State University

Summary

Methods of detecting influential observations for the normal model for censored data are proposed. These methods include one-step deletion methods, deletion of observations and the empirical influence function. Emphasis is placed on assessing the impact that a single observation has on the estimation of coefficients of the model. Functions of the coefficients such as the median lifetime are also considered. Results are compared when applied to two sets of data.

Key words: Censored data; influence function; linear model; one-step methods.

1. Introduction

The detection of influential observations, that is observations whose deletion, either singly or multiply, result in substantial changes in parameter estimates, fitted values or tests of hypothesis, has received considerable attention in recent years. Several methods have been proposed for studying the impact of deletion of observations on parameter estimates obtained from the normal linear model (Belsley, Kuh & Welsch, 1980; Cook & Weisberg, 1982), the logistic regression model (Pregibon, 1981; Johnson, 1985), the Weibull model for censored data (Hall, Rogers & Pregibon, 1982) and the proportional hazards model (Reid & Crépeau, 1985).

The focus of this paper is on the detection of influential observations for the normal regression model fitted to censored data. The methods

Received July 1988; revised December 1988.

[1] Department of Biostatistics, University of Michigan, Ann Arbor, Michigan, USA.

[2] Department of Quantitative Business Analysis, Louisiana State University, Baton Rouge, Louisiana, USA.

discussed are one-step deletion diagnostics, the empirical influence function and deletion of single observations. These methods are applied to the problem of the detection of a single influential observation and to the assessment of the effect that an observation has on parameter estimates and functions of the parameter estimates, such as the median lifetime. The use of these diagnostics is illustrated in several examples.

2. Linear Regression Model for Censored Data

Let T_j denote the failure time of the jth observation. Then the linear regression model takes the form

$$Z_j = \log(T_j) = x_j^T \beta + \sigma e_j \quad (j = 1, \ldots, n) ,$$

where the distribution function F and density function f of the e_js is normal. The covariate vector x_j is p-dimensional with $x_{0j} = 1$ and β is a parameter vector with $\beta = (\beta_0, \ldots, \beta_{p-1})$. The censoring times S_j are independently distributed with distribution function G_j, the failure times Z_j are independently distributed with distribution function F, and $Z_1, \ldots, Z_n$ are independent of $S_1, \ldots, S_n$. We observe

$$Y_j = \min(Z_j, S_j) \quad \text{and} \quad \delta_j = \begin{cases} 1 & \text{if } j\text{th observation is a failure} , \\ 0 & \text{if } j\text{th observation is censored} . \end{cases}$$

This random censorship model includes Type I censoring as a special case.

Two methods for the maximum likelihood estimation of β and σ will be discussed based on Newton–Raphson iteration and the expectation maximization (EM) algorithm. Let

$$u_j(\beta, \sigma) = u_j(\theta) = (Y_j - x_j^T \beta)/\sigma , \tag{1}$$

$\Phi(\cdot)$ denote the standard normal cumulative distribution function, and X be the design matrix. The maximum likelihood estimator of $\theta = (\beta, \sigma)$ is based on $\Lambda(\theta)$, the log-likelihood of the data. In this case

$$\Lambda(\theta) = \sum_j L_j(\theta) ,$$

where $\quad L_j(\theta) = \delta_j \log\left[\phi\{u_j(\theta)\}/\sigma\right] + (1 - \delta_j) \log\left[1 - \Phi\{u_j(\theta)\}\right] .$

The maximum likelihood estimators of β and σ can then be computed using the Newton–Raphson method. For application of this method the following set of equations is solved iteratively:

$$\hat{\theta}_{[k+1]} = \hat{\theta}_{[k]} + I_{[k]}^{-1} q_{[k]} \quad (k = 0, 1, \ldots,) , \tag{2}$$

where

$$q(\theta) = (\partial/\partial\theta)\Lambda(\theta)\,, \quad I(\theta) = -(\partial/\partial\theta)q(\theta) \tag{3}$$

and $I_{[k]}$ and $q_{[k]}$ are $I(\theta)$ and $q(\theta)$ evaluated at $\hat{\theta}_{[k]}$, the estimated value of θ at the kth iteration.

The maximum likelihood estimators of β and σ can also be computed using the EM algorithm (Dempster, Laird & Rubin, 1977; Aitkin, 1981). The expectation step of this algorithm requires the computation of

$$y_j^* = \delta_j y_j + (1-\delta_j)\mathrm{E}(Y_j \mid Y_j > S_j)\,, \tag{4}$$
$$y_j^{*2} = \delta_j y_j^2 + (1-\delta_j)\mathrm{E}(Y_j^2 \mid Y_j > S_j)\,. \tag{5}$$

Then $\hat{\beta}$ and $\hat{\sigma}$ are computed in the maximization step using the ordinary least squares estimates evaluated at y^* and y^{*2}, that is,

$$\hat{\beta} = (X^TX)^{-1}X^Ty^*\,, \tag{6}$$
$$\hat{\sigma}^2 = n^{-1}(y^* - X\beta)^T(y^* - X\beta)\,. \tag{7}$$

For the normal case we obtain

$$\mathrm{E}(Y_j \mid Y_j > S_j) = x_j^T\beta + \sigma\phi(u_j)/\big(1-\Phi(u_j)\big)\,,$$
$$\mathrm{E}(Y_j^2 \mid Y_j > S_j) = (x_j^T\beta)^2 + \sigma^2 + \sigma(S_j + x_j^T\beta)\phi(u_j)/\big(1-\Phi(u_j)\big)\,,$$

where u_j is given by (1). Since y^* is a function of the estimates of β and σ, the EM estimates must be computed iteratively.

Diagnostics for this model can be developed based on either the EM algorithm or the Newton–Raphson method for obtaining estimates; however, the rates of convergence of these two methods differ with the Newton–Raphson method converging quadratically while the EM algorithm converges linearly. Although the convergence rate for the EM algorithm is slower, often it will converge to an estimate when the Newton–Raphson method is divergent.

3. Diagnostics for the Linear Model with Censored Data

We turn now to a discussion of methods that can be used to detect influential observations for a normal model that is fitted to censored data. These methods differ from those used for ordinary linear regression due to the effect of censoring and the iterative procedures required for the calculation of the estimates. We consider four methods for assessing the

influence of a single observation for this model: the empirical influence curve, one-step methods and deletion.

3.1. The Empirical Influence Curve

The influence curve or function can be used to monitor the influence of individual cases on estimates since it can be interpreted as a measure of the change in parameter estimates when the point x is added to the sample. To calculate the influence curve, one considers estimators which are functionals of the empirical cumulative distribution function, $\hat{\theta}_n = T_n(F_n)$, or estimators which can be replaced by functionals asymptotically, that is, $\theta = T(F)$. For a maximum likelihood estimate, the appropriate functional $T(F)$, is the solution to

$$\int q(\theta, y)\, dF(y) = 0 ,$$

where $q(\theta, y) = \partial L(\theta, y)/\partial\theta$. Let $F_\epsilon = (1-\epsilon)F + \epsilon\Delta_y$, where Δ_y is the point mass at y, and $H(\theta, \epsilon) = \int q(\theta, y)\, dF_\epsilon(y)$. Then the influence curve can be obtained by implicit differentiation of the equation $H\{T(F_\epsilon), \epsilon\} = 0$ leading to

$$IC(y; F, T) = I^{-1}(\theta) q(\theta, y_j) . \tag{8}$$

This quantity is the general form of the influence curve for a maximum likelihood estimate. For a linear combination of the elements of β it takes the form

$$IC(y_j, F, R) = \mathbf{Q}\{IC(y_j; F, T)\} ,$$

where $R = \mathbf{Q}\theta$ and $\mathbf{Q}$ is a vector of dimension $p+1$.

The influence curve is commonly estimated by substituting the sample cumulative distribution function $\hat{F}_n$ for F in (8), yielding the empirical influence curve

$$EIC(y_j, \hat{F}, \hat{\theta}) = I^{-1}(\hat{\theta}) q_j^*(\hat{\theta}, y_j) , \tag{9}$$

where $q_j^*(\hat{\theta}, y_j) = \partial L_j(\theta, y_j)/\partial\theta\big|_{\theta=\hat{\theta}}$. This estimator of the influence curve is based on the assumption that an infinitely large sample has been used to obtain $\hat{F}$.

3.2. One-Step Methods

Measures of influence for ordinary least squares are generally based on the change in parameter estimates when the ith observation is deleted, that is, $\hat{\theta} - \hat{\theta}_{(i)}$, where $\hat{\theta}_{(i)}$ is the estimate of θ when the ith observation

is deleted (Cook & Weisberg, 1980; Belsley *et al.*, 1980). This difference measure has been applied in other, more computationally complex settings by using a one-step estimate of $\hat{\theta}_{(i)}$ (Cook & Wang, 1983; Hall *et al.*, 1982) and can be implemented for either the Newton–Raphson iterative method or the EM algorithm. In either case the full data estimate is used as the starting value for the one-step estimate.

The change in estimates, after deletion of the ith observation, based on the one-step Newton–Raphson estimate is given by

$$\Delta NR = \hat{\theta} - \hat{\theta}_{(i)} = I_{(i)}^{-1}(\hat{\theta}) q_i(\hat{\theta}) \,, \tag{10}$$

where q_i is the ith element of $q(\hat{\theta})$ and $I_{(i)}$ is defined by (3) with the ith point removed. It is of interest to note that the one-step Newton–Raphson and empirical influence curve, given by (9), differ only in the use of the ith observation in the computation of $I^{-1}(\hat{\theta})$. The one-step EM algorithm estimate is based on substituting $\hat{\beta}$ and $\hat{\sigma}$ into (4) and (5) to obtain estimators of $y^*_{j(i)}$ and $y^{*2}_{j(i)}$ when the ith point is removed. This yields

$$\begin{aligned} \Delta EM &= \hat{\theta} - \hat{\theta}_{(i)} \\ &= \frac{(X^T X)^{-1} x_i^T [y_i \delta_i - x_i^T \hat{\beta} + (1 - \delta_i)(x_i^T \hat{\beta} + \hat{\sigma} h(\hat{u}_i))]}{1 - x_i^T (X^T X)^{-1} x_i} \end{aligned}$$

where $\hat{u}_i = (y_i - x_i^T \hat{\beta})/\hat{\sigma}$ and $h(\hat{u}_i) = \phi(\hat{u}_i)/(1 - \Phi \hat{u}_i))$.

4. Examples

We consider two examples, Crawford's motorette data and the Stanford heart transplant data. Each of the diagnostics discussed is computed and compared, with interest focusing on the median lifetime in one example and on the slope of the regression line in the other.

4.1. Crawford's Motorette Data

Crawford's (1970) data set considers the failure times of electrical insulation of motorettes as a function of temperature. Ten motorettes were tested at each of the temperatures 220°C, 190°C, 170°C and 150°C, with interest focusing on the median of the distribution at 130°C, which, as is typical in accelerated life tests, was unobserved. The data were modelled by the Arrhenius Law

$$Y_i = \log T_i = \beta_0 + \beta_1 / t_a + \sigma e_i' \qquad (i = 1, \ldots, 40) \,,$$

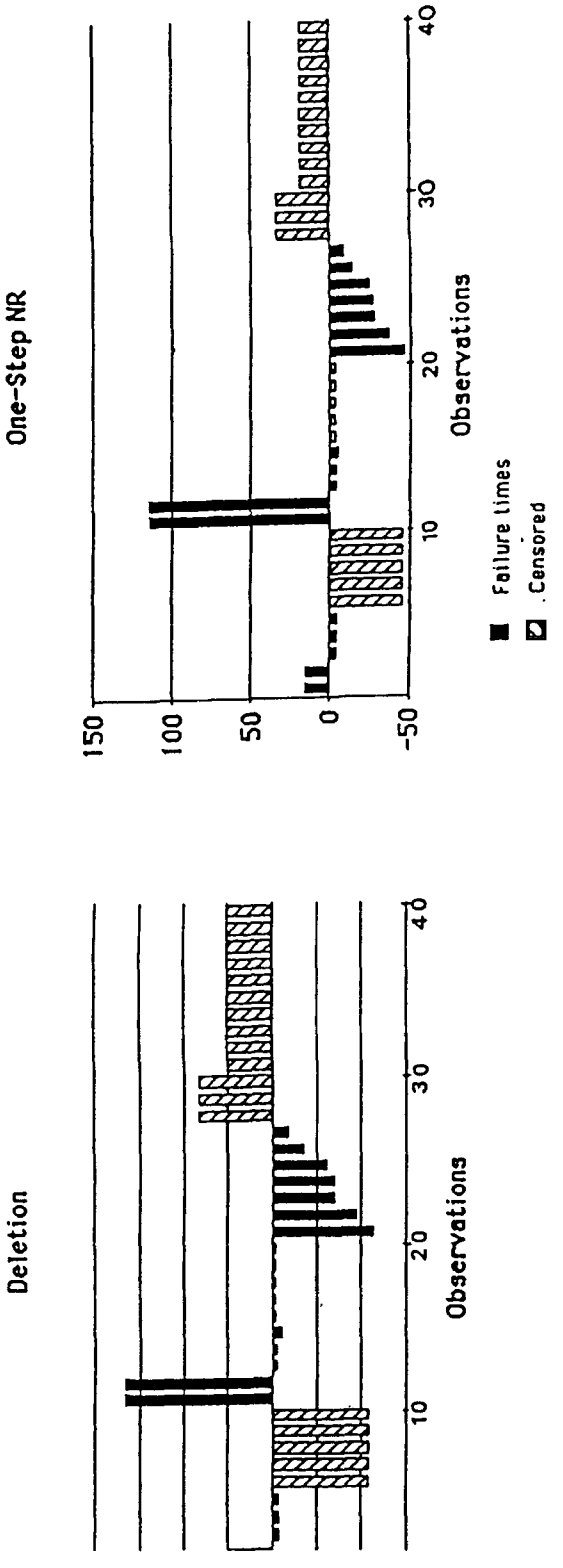

Fig. 1—Influence diagnostics for the median lifetime at 130°C for Crawford's motorette data. (a) Deletion and one-step Newton–Raphson; (b) one-step EM and empirical influence function

where t_a is the absolute temperature and the e_i s are assumed to be normally distributed with mean zero and variance one. The data at 16 months and the following measures of influence for the median lifetime at 130°C are presented in Figures 1a and b: the change in the estimate of median lifetime based on the one-step EM algorithm, the one-step Newton–Raphson given by (10), the empirical influence function (9), and the change in the estimate based on deletion of a single observation. Each of these measures was standardized by dividing the difference by $\hat{\sigma}$. The deletion and one-step Newton–Raphson methods give similar results. These methods isolate the two early failures at 190°C, observations 11 and 12, whereas the influence curve finds these observations less influential than the failures at 170°C, observations 21 and 22, and the one-step EM algorithm does not isolate these values. This result may be due to the linear rate of convergence of the EM algorithm.

It is of interest to note that both influence curve and one-step EM estimate draw greatest attention to the first two failures at 170°C, observations 21 and 22, whereas the one-step Newton–Raphson and deletion methods call greatest attention to the first two failures at 190°C, observations 11 and 12. This may be due to the impact of the two failures at 190°C on the information matrix I. If they have a large impact on I this will be more readily reflected by the deletion and one-step Newton–Raphson methods resulting in a larger value for these diagnostics. In fact, deletion of the failures at 190°C, observations 11 and 12, results in a substantial change in the $\hat{\sigma}$ element of I^{-1} whereas deletion of each failure at 170°C, observations 21 and 22, does not have this effect.

4.2. Stanford Heart Transplant Data

This data set was collected from 184 patients who participated in the Stanford Heart Transplant Study (Miller & Halperin, 1982). $\log_{10}$(survival time) was modelled as a function of age. In this case we are interested in the influence of single observations on the estimate of the coefficient for age. The empirical influence curve, one-step EM algorithm, deletion and one-step Newton–Raphson estimates have been calculated for this example and selected results are presented in Table 1. Once again these measures have been standardized by dividing the difference by $\hat{\sigma}$.

Each of the measures tended to be similar and gave the same ranking for the five observations with the largest values for each diagnostic. The values of these diagnostics ranged from -3.3 to 5.6 for the empirical influence curve, -2.4 to 4.7 for the one-step EM algorithm, -3.3 to 5.9 for deletion and -3.5 to 6.0 for the one-step Newton–Raphson method. The

TABLE 1

Influence diagnostics for the coefficient of age from the model $\log_{10}(\textit{survival time}) = \beta_0 + \beta_1(\textit{age})$ *fitted to the Stanford Heart Transplant Data.* $n = 184$.

Age	Lifetime	δ_i†	Empirical Influence (x10)	Deletion (x10)	One-Step EM (x10)	One-Step NR (x10)
54	1	1	-33	-33	-24	-35
27	22	1	22	22	17	22
28	7	1	25	25	20	26
19	42	1	31	32	24	32
21	1	1	53	54	46	57
12	86	1	35	37	28	37
13	10	1	56	59	47	60
20	5	1	45	46	37	47

† δ_i is a censoring indicator: 0 = censored, 1 = uncensored.

most influential observation, a death at 10 days for a subject aged 13, was found by each of the diagnostics so that in this example all of the methods performed well. It is of interest to note here that the most influential observations are uncensored observations with short lifetimes, indicating that uncensored observations tend to be more influential than censored observations.

5. Summary and Conclusions

Of the methods discussed here both the one-step Newton–Raphson method and deletion tended to call attention to the same set of observations. While the different methods isolated the same set of points for the Stanford heart transplant data, this was not true for the motorette data at 16 months, where the one-step EM algorithm and empirical influence curve differed from deletion and the one-step Newton–Raphson method in choice of most influential observation. The difference may be due to either of the following factors: the impact that the failure time at 190°C has on I^{-1}, or the way in which the two failures at 190°C act together to create a masking effect. While the deletion and one-step Newton–Raphson diagnostics demonstrate a greater impact when either of the failures at 190°C is deleted, this impact is not as likely to be seen with the empirical influence curve or the one-step EM algorithm. Since the empirical influence

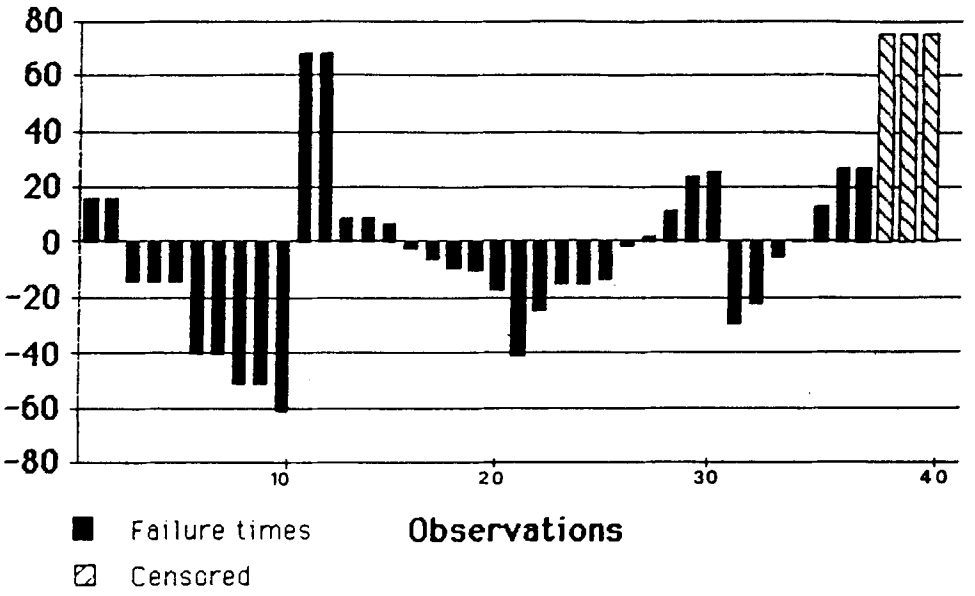

Fig. 2—Influence diagnostics for the median lifetime at 130°C for Crawford's 33 months motorette data.
Deletion result.

curve is measuring the change in parameter estimates when the masking observation is added to the sample, this diagnostic is not as likely to choose the observation as is the one-step Newton–Raphson method. The one-step EM estimate is also unlikely to change much when only one of the two masking points is removed.

These results indicate that the one-step Newton–Raphson diagnostic will perform well; however, information obtained from the empirical influence curve and the one-step EM estimate is helpful in locating observations which are important in the computation of I^{-1}. This does not preclude the use of the EM algorithm for obtaining parameter estimates since these estimates are equivalent to the maximum likelihood estimates and the EM algorithm converges more reliably than the Newton–Raphson method.

One would also expect that uncensored observations are more likely to have an influence on parameter estimates than censored observations, as this is the case for the Kaplan–Meier estimator of the survival function when the influence curve is computed separately for censored and uncensored observations (Reid, 1981). When examining the EM algorithm that is used to obtain parameter estimates one would expect uncensored observations to have a more direct effect on parameter estimates since cen-

sored observations are replaced by their estimated expected values, so that the actual censoring time is used only for obtaining estimates at the first iteration.

To explore this issue in more detail, the 16 month motorette data presented in Figure 1 were examined at 33 months. Figure 2 gives the results obtained with the method of deletion. This data set contains only three censored observations, 38–40, which were tested at 150°C and censored at 17661 hours. These censored observations were the most influential with the deletion method (Figure 2) and also with the one-step EM and empirical influence diagnostics. However, the one-step Newton–Raphson method ranked them second in influence to observations 11 and 12, the first failure at 190°C. The results indicate that censored observations also have an impact on parameter estimates and need to be examined carefully when models are fitted to censored data.

References

AITKIN, M.A. (1981). A note on the regression analysis of censored data. *Technometrics* **23**, 161–163.

BELSLEY, D.A., KUH, E. & WELSCH, R.E. (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity.* New York: Wiley.

COOK, R.D. & WEISBERG, S. (1980). Characterizations of an empirical influence function for detecting influential cases in regression. *Technometrics* **22**, 495–508.

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

COOK, R.D. & WANG, P.C. (1983). Transformations and influential cases in regression. *Technometrics* **25**, 337–343.

CRAWFORD, D.E. (1970). Analysis of incomplete life test data on motorettes. *Insulation/Circuits* **16**, 43–48.

DEMPSTER, A.P., LAIRD, N.M. & RUBIN, D.B. (1977). Maximum likelihood for incomplete data via the EM algorithm (with discussion). *J. Roy. Statist. Soc. Ser. B* **39**, 1–38.

HALL, G.J., ROGERS, W.H. & PREGIBON, D. (1982). Outliers Matter in Survival Analysis. Rand Technical Report D-6761.

JOHNSON, W. (1985). Influence measures for logistic regression: another point of view. *Biometrika* **72**, 59–65.

MILLER, R. & HALPERIN, J. (1982). Regession with censored data. *Biometrika* **69**, 521–531.

REID, N. (1981). Influence functions for censored data. *Ann. Statist.* **9**, 78–92.

REID, N. & CRÉPEAU, H. (1985). Influence functions for proportional hazards regression. *Biometrika* **72**, 1–9.