Statistics & Probability Letters 9 (1990) 67-73
North-Holland

January 1990

INFLUENCE DIAGNOSTICS FOR THE WEIBULL MODEL FIT TO CENSORED DATA

Lisa A. WEISSFELD

Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, USA

Helmut SCHNEIDER

Department of Quantitative Business Analysis, Louisiana State University, Baton Rouge, LA 70803, USA

Received September 1988
Revised December 1988

Abstract: Methods for detecting influential observations for the Weibull model fit to censored data are discussed. These methods include: one-step deletion diagnostics, influence functions and curvature diagnostics. Results indicate that the curvature diagnostics may be helpful in detecting masking.

Keywords: Generalized linear model, maximum curvature, one-step estimates, Weibull model.

1. Introduction

The detection of influential observations, that is, observations whose deletion result in substantial changes in parameter estimates or functions of the parameter estimates, is of great importance when models are fit to censored data. This problem has been considered for both parametric models (Hall, Rogers and Pregibon, 1982; Weissfeld and Schneider, 1988) and the semiparametric proportional hazards model (Reid and Crépeau, 1985). It has been approached from several different directions such as deletion of observations and perturbation of data points.

The focus here is on diagnostics for the Weibull model, although these results can also be applied to other parametric models for censored data such as the gamma, log-logistic, log-normal and exponential models. Diagnostics are developed for examining the effect of a single observation on parameter estimates and estimates of the pth percentile of the survival distribution. The influence curve and a one-step estimate based on fitting the Weibull model as a generalized linear model (GLM) are computed and compared with both the one-step estimates discussed by Hall, Rogers and Pregibon (1982) and deletion of the observation.

The other approach to this problem, namely, examining the effect that perturbation of the data has on parameter estimates is also considered. The maximum curvature of the influence graph and the eigenvector corresponding to the maximum curvature are computed and examined in this case (Cook, 1986). This method has the advantage of isolating points that may have a 'masking' effect, that is, points which effect parameter estimates jointly and will not be detected by single case deletion diagnostics. These 'masking' points act as a group and are often not isolated through the use of single case deletion techniques.

Escobar and Meeker (1987) describe how to use the local influence methods, introduced by Cook (1986), to detect data/model perturbations that have important effects on maximum likelihood estimates of regression model parameters and functions of these parameters, based on arbitrarily censored data. They use perturbations to case weights, observed responses, and the assumed value of the distribution

0167-7152/90/$3.50 © 1990, Elsevier Science Publishers B.V. (North-Holland)

shape parameter to study influence on the full parameter vector, single parameters, and distribution percentiles. Their results apply to location-scale distributions, but their actual applications are with the Weibull and log-normal accelerated failure time models. Escobar and Meeker (1988) provide SAS macros which can be used for the local influence analysis and produce related graphical output. Meeker and Escobar (1988) explore the graphical output and explore the relationship between local and global influence and show how to use local influence analysis to guide the more complicated and computationally intensive global influence analysis.

2. Weibull model for censored data

Let T_j denote the failure time of the jth observation, x_j a p-dimensional covariate vector with $x_{0j} = 1$ and β a parameter vector with $\beta = (\beta_0, \ldots, \beta_{p-1})$. Then, if T_j is distributed according to the Weibull distribution, $\log(T_j)$ will follow an extreme value distribution and the failure times, can be modelled as

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

where e_j follows an extreme value distribution. The censoring times (S_j) are assumed to be independently distributed with distribution function G, so that

$$Y_j = \min(A_j,\ S_j)$$

and $\delta_j = I_{\{Y_j = A_j\}}$ are the observed random variables of interest. It is also assumed that A_j is independent of S_j so that the censoring is random.

Parameter estimates of β and σ can be obtained using maximum likelihood estimation with score vector $U(\beta, \sigma)$ containing components of the form:

$$\frac{\partial \log L}{\partial \beta_j} = \sigma^{-1} \sum_{i=1}^{n} \left(-\delta_i x_{ij} + x_{ij} \exp(z_i)\right), \tag{1}$$

$$\frac{\partial \log L}{\partial \sigma} = -\sigma^{-1} \sum_{i=1}^{n} \left(\delta_i + \delta_i z_i - z_i \exp(z_i)\right), \tag{2}$$

where $z_i = (y_i - x_i^{\mathrm{T}}\beta)/\sigma$. The information matrix takes the form

$$I_{ij}(\beta, \sigma) = \frac{\partial^2 \log L}{\partial \beta_k \partial \beta_j} = -\sigma^{-2} \sum_{i=1}^{n} x_{ik} x_{ij}\, \mathrm{e}^{z_i}, \quad k,\ j = 1, \ldots, p, \tag{3}$$

$$I_{p+1,p+1}(\beta, \sigma) = \frac{\partial^2 \log L}{\partial \sigma^2} = \sigma^{-2} \sum_{i=1}^{n} \left(\delta_i + 2\delta_i z_i - \left(2z_i + z_i^2\right) \mathrm{e}^{z_i}\right), \tag{4}$$

$$I_{j,p+1}(\beta, \sigma) = \frac{\partial^2 \log L}{\partial \beta_j \partial \sigma} = \sigma^{-2} \sum_{i=1}^{n} \left(\delta_i x_{ij} - (x_{ij} + x_{ij} z_i)\, \mathrm{e}^{z_i}\right), \tag{5}$$

with the asymptotic covariance matrix estimated by $I(\beta, \sigma)^{-1}$ evaluated at $(\hat{\beta}, \hat{\sigma})$. The solution to the likelihood equations is obtained using a numerical iterative method such as the Newton–Raphson method. For application of this method, the following set of equations is solved iteratively:

$$\hat{\theta}_{[k+1]} = \hat{\theta}_{[k]} + I_{[k]}^{-1} U_{[k]} \quad (k = 0, 1, \ldots), \tag{6}$$

where $\theta = (\beta, \sigma)$ and I^{-1} and U are evaluated at $\theta_{[k]}$.

An alternative method for obtaining parameter estimates of β and σ is based on the generalized linear model approach (McCullagh and Nelder, 1983; Aitkin and Clayton, 1980). In this case the censoring indicator δ_i is assumed to follow a Poisson distribution with mean

$$\mu_i = t_i^{\sigma} \exp\left(x_i^{\mathrm{T}}\beta\right), \quad i = 1, \dots, n. \tag{7}$$

Parameter estimates are obtained by fitting the log-linear model in GLIM with offset σ log t_i and computing $\hat{\sigma}$ from

$$\sum_{i=1}^{n} \delta_i/\sigma = \sum_{i=1}^{n} \left(t_i^{\hat{\sigma}} \exp\left(x_i^{\mathrm{T}}\hat{\beta}\right) - \delta_i\right) \ln t_i. \tag{8}$$

The estimation begins with $\hat{\sigma} = 1$, fits the log-linear model with offset σ log t_i then uses equation (8) to obtain a new estimate of $\hat{\sigma}$. The procedure oscillates between the last two steps until convergence with suitable damping of $\hat{\sigma}$.

3. Influence diagnostics for parameter estimates

The impact of a single observation on parameter estimates can be ascertained using either deletion diagnostics or the empirical influence curve. Deletion diagnostics, where the parameter estimate is computed with the ith observation deleted from the data set, can be readily computed for an ordinary linear regression analysis; however, since iterative methods are used to obtain parameter estimates for the Weibull model, it is necessary to use one-step approximations based on either the Newton–Raphson method or the generalized linear model approach. For the Newton–Raphson method the one-step approximation to the deleted value of $\theta = (\beta, \sigma)$ is given by

$$\hat{\theta}_{(i)} = \hat{\theta} + I_{(i)}^{-1}(\hat{\theta}) U_{(i)}(\hat{\theta}), \tag{9}$$

where $U_{(i)}$ and $I_{(i)}$ are defined by (6) with the ith point deleted (Hall, Rogers and Pregibon, 1982).

The one-step GLM estimate is obtained by computing the one-step Newton–Raphson estimate of β, obtained from equation (1), with the values of $\hat{\beta}$ and $\hat{\sigma}$ based on the full data set as starting values. A one-step estimate of $\hat{\sigma}$ is then computed using (8).

The impact of a single observation on parameter estimates can also be assessed by computing the influence curve. This quantity measures the effect that the addition of the point x to the sample has on parameter estimates and functions of parameter estimates. Let $\hat{\theta}_n = T_n(F_n)$, where F_n is the empirical cumulative distribution function, so that $\hat{\theta}$ can be written as a functional of the empirical distribution function or can be replaced by a functional asymptotically, that is $\theta = T(F)$. The general form of the influence curve of a maximum likelihood estimate is given by

$$\mathrm{IC}(y_j, F, T) = I^{-1}(\theta) U_j(\theta, y_j),$$

where u_j is the jth term of the sum used to compute the score function.

This result can also be extended to linear combinations of the elements of θ. If $R = Q\theta$, where Q is a vector of dimension $p + 1$, then the influence curve for R is

$$\mathrm{IC}(y_j, F, R) = Q\left\{\mathrm{IC}(y_j, F, T)\right\}.$$

This result can be applied to compute the influence curve for the pth percentile of the extreme value distribution which is given by $y_p = x\beta + \sigma \log(-\log(1 - p))$. The most commonly used estimate of this curve is the empirical influence curve which takes the form

$$\mathrm{EIC}\left(y_j, \hat{F}, \hat{\theta}\right) = I^{-1}(\hat{\theta}) U_j^{*}, \tag{10}$$

where $U_j^{*}(\hat{\theta}, y_j)$ is the jth term of the sums composing the score vector.

Influence can also be assessed by the use of measures of curvature (Cook, 1986). These measures can be obtained by examining the effect of perturbations, in either the elements of the score vector or the covariates, on the parameter estimates obtained from the model. If we let the vector $\omega = (\omega_1, \ldots, \omega_n)$ denote the vector of perturbations, then the curvature is defined by

$$C_l = 2\,|\,l^{\mathrm{T}}\ddot{F}l\,|$$

where $|\,l\,| = 1$, the $(i,\ j)$th element of $\ddot{F}$ is given by $\partial^2 L(\hat{\theta}_\omega)/\partial\omega_i\partial\omega_j$ and $\hat{\theta}_\omega$ is the estimate of θ based on the perturbed data. The matrix $\ddot{F}$ can be more easily computed through the use of the relation

$$\ddot{F} = \Delta^{\mathrm{T}} I^{-1} \Delta.$$

where the $(i,\ j)$th element of Δ is given by $\partial^2 L(\theta\,|\,\omega)/\partial\theta_i\partial\omega_j$ evaluated at $\theta = \hat{\theta}$, $\omega = \omega_0$ and $L(\theta\,|\,\omega)$ is the perturbed likelihood. In the case of the Weibull model, one possible perturbation scheme involves perturbation of each element of the sum comprising the likelihood function, so that the likelihood is of the form

$$L(\theta\,|\,\omega) = \sum \omega_i \log L_i(\theta)$$

and

$$\Delta_{ij} = \begin{cases} -\delta_i x_{ij} + x_{ij}\exp(z_j), & i = 1, \ldots, p,\ \ j = 1, \ldots, n, \\ -\dfrac{\delta_j}{\sigma} - \delta_j \dfrac{z_j}{\sigma} + \dfrac{z_j}{\sigma}\exp(z_j), & i = p+1,\ \ j = 1, \ldots, n. \end{cases} \tag{11}$$

Thus C_l and $\ddot{F}$ can be easily computed. The maximum curvature, $C_{\max}$, is the maximum eigenvalue of $\ddot{F}$ and the corresponding eigenvector is given by $l_{\max}$. Elements of $l_{\max}$ are then examined individually with large values pointing to observations which are possibly influential.

The effect of perturbation of the independent variables can also be examined using this approach (Cook, 1986). For example, if the lth covariate is perturbed by amount ω then

$$x'_{lj} = x_{lj} + \omega_j s_j \tag{12}$$

defines the perturbed covariate, where s_j is a scale factor. The maximum curvature of the influence graph is computed as above with Δ being defined as a $(p+1) \times (p+1)n$ matrix. If we partition this matrix into $(p+1)$ submatrices, so that $\Delta = (\Delta_1, \ldots, \Delta_{p+1})$, then the lth submatrix is defined by

$$\Delta_{l_{jk}} = \begin{cases} -\dfrac{s_l x_{kj}\hat{\beta}_l}{\hat{\sigma}}\exp(z_j), & k \neq l,\ k = 1, \ldots, p,\ \ j = 1, \ldots, n, \\ -\dfrac{s_l\delta_j}{\hat{\sigma}} + \dfrac{s_l\left(1 - x_{lj}\hat{\beta}_l\right)}{\hat{\sigma}}\exp(z_j), & k = l,\ \ j = 1, \ldots, n, \\ \left(\hat{\beta}_l\left(\delta_j - \exp(z_j)\right) - \dfrac{z_j\hat{\beta}_l}{\hat{\sigma}^2}\exp(z_j)\right)s_l, & j = 1, \ldots, n,\ \ k = p+1, \end{cases} \tag{13}$$

where $s = (s_1, \ldots, s_{p+1})$ is a scale factor used to account for the different measurement units associated with the covariate vector x. This diagnostic indicates the effects of data perturbation on the lth coefficient in the regression model by examining $l_{\max}$, the eigenvector corresponding to the maximum curvature.

4. Examples

In order to illustrate the use of these methods several sets of data will be examined. Table 1 presents influence diagnostics for Crawford's (1970) motorette data. This data examines the influence of individual

Table 1
Influence diagnostics for log median lifetime at 130 ° C and curvature diagnostics for time and temperature for Crawford's motorette data

Temperature ° C	Lifetime (hours)	δ_i	No. of observations	One-step GLM	Empirical influence	One-step NR	Deletion	Curvature (times)	Curvature (temp.)
220	408	1	2	0.0120	0.0027	0.0029	0.0030	−0.0594	−0.0187
220	504	1	3	0.0059	−0.0079	−0.0079	−0.0084	−0.1286	−0.1204
220	600	1	2	−0.0042	−0.0176	−0.0180	−0.0191	−0.1508	−0.2740
220	648	1	2	−0.0108	−0.0221	−0.0230	−0.0240	−0.1417	−0.3733
220	696	1	1	−0.0183	−0.0261	−0.0280	−0.0286	−0.1174	−0.4892
190	408	1	2	−0.0385	0.0082	0.0146	0.0096	0.5242	0.0774
190	1344	1	2	−0.0173	−0.0154	−0.0153	−0.0157	0.0829	0.0659
190	1440	1	1	−0.0155	−0.0162	−0.0161	−0.0167	0.0555	0.0613
190	1920	1	1	−0.0070	−0.0169	−0.0166	−0.0176	−0.0545	0.0225
190	2256	1	1	−0.0007	−0.0139	−0.0136	−0.0145	−0.1050	−0.0240
190	2352	1	1	0.0012	−0.0125	−0.0122	−0.0131	−0.1152	−0.0407
190	2596	1	1	0.0062	−0.0077	−0.0075	−0.0081	−0.1132	−0.0909
190	3360	1	1	0.0460	0.0562	0.0703	−0.0583	0.0781	−0.4067
170	1764	1	1	−0.0545	−0.0285	−0.0283	−0.0275	0.3612	0.0788
170	2772	1	1	−0.0457	−0.0363	−0.0367	0.0369	0.1957	0.0788
170	3444	1	1	−0.0402	−0.0386	−0.0388	−0.0471	0.1129	0.0759
170	3542	1	1	−0.0394	−0.0387	−0.0389	−0.0403	0.1021	0.0752
170	3780	1	1	−0.0375	−0.0389	−0.0390	−0.0407	0.0770	0.0732
170	4680	1	1	−0.0295	−0.0375	−0.0373	−0.0398	−0.0041	0.0606
170	5196	1	1	−0.0245	−0.0351	−0.0349	−0.0374	−0.0417	0.0491
170	6206	1	1	−0.0133	−0.0269	−0.0169	−0.0190	−0.0970	0.0158
170	7716	0	1	0.0570	0.0553	0.0624	0.0568	−0.0152	−0.1432
170	7884	0	1	0.0605	0.0590	0.0673	0.0607	−0.0080	−0.1553
150	11781	0	1	0.0166	0.0135	0.0140	0.0138	−0.0409	0.0005
150	12453	0	1	0.0191	0.0158	0.0164	0.0163	−0.0428	−0.0015
150	13897	0	1	0.0253	0.0216	0.0227	0.0223	−0.0454	−0.0076
150	14469	0	1	0.0280	0.0243	0.0256	0.0252	−0.0458	−0.0107
150	15891	0	1	0.0358	0.0317	0.0339	0.0330	−0.0447	−0.0202
150	17325	0	2	0.0452	0.0404	0.0443	0.0424	−0.0403	−0.0329
150	17661	0	3	0.0477	0.0427	0.0471	0.0448	−0.0387	−0.0364

* $\delta_i = 1$ if observation is uncensored and 0 if observation is censored.

observations on the estimated median lifetime when the motorette is subjected to a temperature of 130 ° C. The data set consists of 40 observations with 10 motorettes tested at each of the following temperatures: 220 ° C, 190 ° C, 170 ° C and 150 ° C. The Arrhenius law was used to model this data with the model

$$Y_i = b_0 + b_1/t_{ai} + \varepsilon_i, \quad i = 1, \ldots, 40,$$

where t_{ai} is the absolute temperature of the ith observation. The results given in the table are based on the difference in the estimated median lifetime at 130 ° C when the ith observation is deleted from the data set. Each of the methods is looked at; the one-step GLM estimate, the empirical influence curve, the one-step Newton–Raphson estimate and the change based on deletion of the observation from the data set.

The last two columns of Table 1 list the results obtained using the curvature diagnostic based on (11) and (13). The first column presents the results obtained when each element in the sum of the likelihood is perturbed. These results indicate that the two early failure times of 408 at 190 ° C are highly influential. Due to the masking effect of these observations, this result would not be obtained from single case deletion diagnostics since one of these points remain in each computation. In fact, when these points are deleted from the data set, the coefficient associated with temperature changes by approximately on third of one standard deviation and the estimated median lifetime at 130 ° C becomes 55207 hours or 1207 hours shorter than the estimated median lifetime of 56414 hours which is based on the whole data set. The first failure time at 170 ° C is also flagged through the use of this diagnostic with deletion resulting in an estimated median lifetime that is 1571 hours longer than the estimate based on the whole data set.

The last column present results that are obtained when the temperature values are perturbed, as defined by (12). In this case the largest failure times at 220 ° C and 190 ° C are flagged rather than the smallest ones as was the case when individual elements in the likelihood are perturbed. Deletion of the largest failure time at 220 ° C causes the estimated median lifetime at 130 ° C to be 1649 hours longer than the estimate based on the whole data set, while deletion of the largest failure time at 190 ° C causes the estimated median lifetime at 130 ° C to be 3193 hours shorter than the estimate based on the whole data set.

Comparison of these methods in this example illustrate the differences obtained when using these diagnostics. In general, the deletion methods flagged the same set of observations with the exception of the one-step GLM method. The one-step GLM estimate flags the first failure time at 170 ° C as being influential; whereas the other methods do not flag this observation. The relative ordering of observations flagged by the GLM method is also different. While the failure time of 1764 at a stress temperature of 170 ° C is flagged as being the first or second most influential observation by the other deletion methods, it is flagged as being the fifth most influential by the one-step GLM estimate. It is also of interest to note that the order of importance of the first and second observations is reversed when comparing the Newton–Raphson method to the results obtained from deletion or the empirical influence curve. In this example the empirical influence curve results matched those obtained from deletion so that this method may be preferable in this instance. The importance of the curvature diagnostics for flagging points which are not flagged with single-case deletion diagnostics is illustrated by the 'masking' points which occur at the first failure time at 190 ° C.

Table 2 presents results obtained from fitting a Weibull model to the epoxy stress data from Andrews and Herzberg (1986). The data consist of 108 vessels which were tested at different levels of stress ranging from 68% to 86%. The model fit was

$$Y_i = \beta_0 + \beta_1 x_i + \sigma e_i, \quad i = 1, \ldots, 108,$$

where x_i denotes the level of stress. Interest is focused on the estimate of median lifetime at 50% stress.

The results from this analysis indicate that each of the diagnostics yield very similar results with the one-step GLM and one-step NR being quite similar. These diagnostics flag the large failure times at the largest and smallest stress levels as being influential which is to be expected since large or small failure

Table 2
Influence diagnostics for log median lifetime at 50% level of stress and curvature diagnostics for time and stress for epoxy vessel data

Stress	Time	δ_i*	One-step GLM	Empirical influence	One-step NR	Deletion	Curvature (times)	Curvature (stress)
68	4000	1	−0.1652	−0.1741	−0.1745	−0.1794	0.0499	−0.0311
68	5376	1	−0.1562	−0.1703	−0.1708	−0.1758	0.0491	−0.0321
68	7320	1	−0.1450	−0.1640	−0.1648	−0.1698	0.0475	−0.0333
86	7552	1	−0.1261	−0.1259	−0.1304	−0.1307	0.2308	0.2665
68	8616	1	−0.1381	−0.1596	−0.1606	−0.1654	0.0464	−0.0340
86	1108.2	1	−0.1499	−0.1367	−0.1427	−0.1422	0.2645	0.3000
86	1148.5	1	−0.1559	−0.1392	−0.1456	−0.1450	0.2729	0.3084
86	1569.3	1	−0.2171	−0.1614	−0.1724	−0.1690	0.3546	0.3917
86	1750.6	1	−0.2426	−0.1690	−0.1823	−0.1775	0.3871	0.4255
86	1802.1	1	−0.2498	−0.1710	−0.1850	−0.1797	0.3960	0.4349
68	9120	1	−0.1356	−0.1578	−0.1589	−0.1637	0.0459	−0.0342

* $\delta_i = 1$ if observation is uncensored and 0 if observation is censored.

times tend to be most influential in this setting. This may be due to the fact that few observations are censored in this data set and the censoring is Type I.

Results obtained from the use of the curvature diagnostics indicate that the larger failure times at the largest stress level were most influential. The small failure times were not flagged as being influential. In this case the diagnostics yielded the same results when using perturbation of an element in the likelihood and perturbation of the stress level. These results do indicate that the large failure times do not have the same impact that they do in the case of the deletion diagnostics, since the deletion diagnostics flagged the large failure times at the two extreme levels of stress, while the curvature diagnostics flagged the largest failure times in the data set. Although 16 of the vessels subjected to a 68% level of stress were censored at 9973 hours, this set of observations was not flagged as influential by any of the methods considered.

5. Summary and conclusions

These examples illustrate the usefulness of curvature diagnostics and one-step diagnostics in locating influential points in data analysis. The curvature diagnostics have the advantage of pointing out some cases of masking that will go undetected with standard deletion diagnostics. In the two examples considered both large and small observed times had an impact on parameter estimates and functions of parameter estimates. Censored observations also tended to be less influential than uncensored observations in each of the two examples that were considered.

References

Aitkin, M. and D. Clayton (1980), The fitting of exponential, Weibull and extreme value distributions to complex censored survival data using GLIM, *Appl. Statist.* **29**, 156–63.

Andrews, D.F. and A.M. Herzberg (1986), *Data: A Collection of Problems from many Fields for the Student and Research Worker* (Springer, New York).

Cook, R.D. (1986), Assessment of local influence, *J. Roy. Statist. Soc.* **48**, 133–155.

Escobar, L.A. and W.Q. Meeker (1987), Assessing local influence in regression analysis with censored data, Paper presented at teh 147th Annual Meeting of the American Statistical Association, San Francisco, CA, August 1987.

Escobar, L.A. and W.Q. Meeker (1988), Using the SAS system to assess local influence in regression analysis with censored data, *Proc. Annual SAS User's Group International Conference.*

Hall, G.J., W.H. Rogers and D. Pregibon (1982), Outliers matter in survival analysis, Rand Technical Report D-6761.

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

Meeker, W.Q. and L.A. Escobar (1988), Influence diagnostics for reliability data, Paper presented at the 148th Annual Meeting of the American Statistical Association, New Orleans, LA, August 1988.

Reid, N. and H. Crépeau (1985), Influence functions for proportional hazards regression, *Biometrika* **72**, 1–10.

Weissfeld, L.A. and H. Schneider (1988), Influence diagnostics for the normal linear model with censored data, *Austral. J. Statist.*, to appear.