Using the SAS* System to Assess Local Influence in Regression Analysis with Censored Data

Luis A. Escobar, Dept. of Experimental Statistics, Louisiana State University
William Q. Meeker, Jr., Dept. of Statistics, Iowa State University

Abstract
This paper describes a SAS* macro for the assessing local influence in regression analysis with censored data. The macro allows one to identify those observations that have the greatest local influence on the estimates of the parameters or functions of the parameters of one's model.

1 Introduction
Inferences in survival data analysis are frequently based on a parametric regression model. Sometimes we have empirical evidence or subject matter knowledge that the assumed model will provide an adequate approximation to make inferences. Most commonly, we have very little information on the validity of the assumptions and we need to assess the impact of possible model departures on the inferences. This can be done using "global" and "local" measures of influence. Cook (1986) discusses the differences between these two types of influence measures and develops general theory for local influence analysis. Escobar and Meeker (1988) give detailed information on the application of Cook's methodology to regression analysis with arbitrarily censored data.

Case deletion is a popular method of global influence for linear regression (see for example, Beasley, Kuh, and Welsch 1980 and Cook and Weisberg 1982). Some of these ideas have also been explored in logistic and Weibull regression models (see Pregibon 1981 and Hall, Pregibon and Rogers 1982). Outside of linear regression, global influence analysis has the disadvantages of being computationally intensive and difficult to summarize and interpret, due to the large amount of information that is generated.

Local influence statistics provide scientists with a systematic method to study and compare the effects that small perturbations of the model will have on the inferences of interest. This is useful in the identification of model departures that have a strong influence on relevant inferences. Here we show how to use SAS* to implement the general method of Cook (1986) when applied to the analysis of censored regression data.

2 Model and Model Estimation
2.1 Regression Model
We assume that there are n independent lifetime observations \( t_1, t_2, \ldots, t_n \), which are either failures times or right censored observations (see Nelson 1982 for definitions of types of censoring).

We assume that \( t_i \) is an observation of a random variable \( T_i \), such that \( Y_i = \log(T_i) \) follows a location-scale distribution, \( G \), with location parameter \( \mu_i \) and constant scale parameter \( \sigma \). We also assume that \( \mu_i = \xi'_i \beta \), where \( \xi_i = (\xi_{i1}, \xi_{i2}, \ldots, \xi_{ip})' \) is a vector of explanatory variables and \( \beta = (\beta_1, \ldots, \beta_p)' \) is a vector of unknown regression coefficients. These assumptions imply that the cdf and pdf functions of \( Y_i \) are given by \( G(t_i) = \Phi(\xi_i) \) and \( g(t_i) = (1/\sigma) \phi(\xi_i) \), where \( \xi_i = (y_i - \mu_i)/\sigma \), \( y_i = \log(t_i) \), and \( \Phi(\cdot) \) and \( \phi(\cdot) \) are the standardized forms of the location-scale distribution (i.e. when \( \mu_i = 0 \) and \( \sigma = 1 \)).

The unknown parameters \( \theta = (\beta', \sigma)' \) are defined in a parameter space \( \Theta \) and they will be estimated from the data.

2.2 Log-likelihood
The log-likelihood under the assumed model is

\[
L(\theta) = \sum_{i=1}^n L_i
\]

where \( L_i = \log(1 - \Phi(\xi_i)) \) if \( t_i \) is a right censored observation and \( L_i = \log(\phi(\xi_i)/\sigma) \) if \( t_i \) is a failure time.

The parameters \( \theta \) are estimated using maximum likelihood. It is assumed that, for any data/model combination, \( L(\theta) \) has a unique maximum that is not located on the boundary of \( \Theta \).

We restrict the discussion to data sets containing failure times and right censored observations because these are the only two types of observations that are currently handled by SAS*. Fortunately, this is all that is needed for most practical applications. For more details, the derivation of the model and its log-likelihood, and the generalization to arbitrarily censored data, see Kalbfleisch and Prentice (1980), Lawless (1982, Chapter 6), or Escobar and Meeker (1988).

3 Local Influence—Model Perturbations
The local influence analysis allows one to assess the effects that small changes in the assumed model will have on inferences. Our approach here is based on the study of the log-likelihood displacement as proposed in Cook (1986).

3.1 Log-likelihood Displacement
We entertain model perturbations that are determined by \( q \)-dimensional vectors, \( \omega \), located in a certain space \( \Omega \). Every \( \omega \) in \( \Omega \) determines a direction and a magnitude of model perturbation. We assume that \( \Omega \) contains a point \( \omega_0 \) that coincides with the assumed model. Let \( L(\theta | \omega) \) be the log-likelihood of the perturbed model corresponding to a point

1036
in \( \mathbf{f} \). Then a basic quantity of interest is the log-likelihood displacement,

\[
LD(\omega) = 2[L(\hat{\theta}) - L(\hat{\theta}_0)]
\]

where \( \hat{\theta} \) and \( \hat{\theta}_0 \) are the MLE estimators of \( \theta \) for the assumed and the perturbed model, respectively. Observe that \( LD(\omega) \geq 0 \) and \( LD(\omega_0) = 0 \).

In a sense \( LD(\omega) \) measures the decrease in maximum log-likelihood that results from moving away from \( \omega_0 \) to \( \omega \). If \( \omega_0 \) is near \( \omega \) and \( LD(\omega) \) is large, then the inferences on the parameter vector \( \hat{\theta} \) are not robust to small model departures and the model, the data, or the data/model combination must be carefully scrutinized. For a detailed description of the theory, computation, and interpretation of these diagnostics, see Cook (1986) and Escobar and Meeker (1988).

3.2 Assessment of Local Influence

When the entire parameter vector \( \theta \) is of primary interest, Cook (1986) proposes that one assess local influence by determining the maximum curvature vector, \( h_{\max} \), and the corresponding direction of maximum curvature vector, \( h_{\text{max}} \). We concentrate on \( h_{\max} \), which is invariant to reparameterizations of the model.

To compute \( h_{\max} \) define the \( q \times q \) matrix

\[
F = -\Delta Q^{-1} \Delta
\]

where \( \Delta \) is the \( q \times q \) matrix with elements

\[
\Delta_{ij} = \frac{\partial^2 L(\theta | \omega)}{\partial \theta_i \partial \theta_j} (\text{evaluated at } (\hat{\theta}, \omega_0))
\]

and \( Q \) is the observed information matrix for the postulated model

\[
Q_{ij} = \frac{\partial^2 L(\theta)}{\partial \theta_i \partial \theta_j} (\text{evaluated at } \hat{\theta}).
\]

The direction of maximum curvature vector \( h_{\max} \) is the eigenvector corresponding to the largest (in absolute value) eigenvalue of \( F \) in (1).

When the inferential interest centers on a subset of the components of \( \theta \) rather than on the whole vector of parameters, the definition of the log-likelihood displacement is modified as described in Cook (1986) and Escobar and Meeker (1988). Without loss of generality, assume that \( \theta = (\theta', \theta'') \) and that inferences are desired primarily on \( \theta' \). Redefine the matrix \( F \) by

\[
F = -\Delta' Q^{-1} - B_{22} \Delta
\]

where

\[
B_{22} = \begin{bmatrix} 0 & 0 \\ 0 & Q_{22}^{-1} \end{bmatrix}
\]

and \( Q = \begin{bmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{bmatrix} \).

The scale parameters are nonhomogeneous and defined by

\[
s_i(\omega) = \sigma \exp(\omega_i)
\]

and \( \omega_0 = (0, \ldots, 0)' \).

3.3 Model Perturbations

Currently the macro LOCINF provides local influence analysis for the following model perturbations.

a) "Case weight perturbations"

Here the perturbed log-likelihood is defined by

\[
L(\theta | \omega) = \sum_{i=1}^n \omega_i L_i
\]

and \( \omega_0 = (1, \ldots, 1)' \).

b) "Distribution shape perturbation"

The scale parameters are nonhomogeneous and defined by

\[
s_i(\omega) = \sigma \exp(\omega_i)
\]

and \( \omega_0 = (0, \ldots, 0)' \).

c) "Response times perturbation"

The observed \( \log(\text{lifetimes}) \) are perturbed by

\[
y_i(\omega) = y_i + \omega_i z_i
\]

where \( s \) is a scale factor and \( \omega_0 = (0, \ldots, 0)' \).

Escobar and Meeker (1988) give the derivatives to compute the elements in (2) for the above model perturbations, arbitrary censoring, and a general location-scale distribution.

4 Macro LOCINF

In this section we describe the current capabilities of the macro LOCINF for the assessment of local influence. This macro is still being developed and we will be adding new features to enhance its performance.

4.1 Parameters

The macro LOCINF has the following parameters:

SURDATA name of the data set containing the survival data to be analyzed.

T the name of the failure time variable \( T \). This must be one of the variables in SURDATA.

X the names of the explanatory variables. These is a list of valid variables names separated by blanks and they must be defined in SURDATA.

CVAR the name of the variable in SURDATA containing the censoring codes.

CCODE the value of the CVAR variable that indicates right censored observations. The default value is equal to 1.

QUADATA name of the data set containing the definition of the quantiles for which a influence analysis is requested. This data set must contain the variables listed in \( X \).

P the name of a variable in QUADATA that contains the probability value that defines the \( p \) quantile at the corresponding levels of the \( X \) variables.
DIST the name of the failure time distribution. The default value is WEIBULL; the other possible choice is LNORMAL. Using an exponentiation transformation this also allows the analysis of normal and smallest extreme value data.

PERT the perturbation to be used in the influence analysis. The default value is equal to WEIGHT; other choices are RESPONSE and SHAPE.

RRPLOTS an indicator of whether or not regression and residual plots are to be drawn. The default value is YES; the other possibility is NO.

CASEID an indicator of whether or not to identify the observations in the residual and regression plots. The default value is NO.

ALPHA a number in [0.5,1) used in defining the quantiles \( t(\alpha) \) and \( t(1-\alpha) \) that are drawn in the scatter and the residual plots. The default values is equal to 0.05.

HPLOT an indicator of whether or not the vectors of maximum normal curvature are to be drawn. The default values is equal to YES; the other possibility is NO.

GOODOUT the name of an optional output data set that contains all the information used to generate the plots for goodness of fit.

DIROUT the name of an optional output data set containing the direction of the vectors of maximum normal curvature.

QUAOUT the name of an optional output data set that contains the estimated values of all the quantiles specified in QUADATA.

4.2 Output

The macro provides three different types of information.

a) Goodness of Fit Plots: These plots are not by themselves part of the local influence approach but they are very useful in interpreting the local influence diagnostics. Also any influence analysis is meaningless unless that the assumed model is a reasonable descriptor of the data.

- Scatter plot of the data versus the explanatory variable when there is only one explanatory variable in the model. The plot includes the original data \( t_i \), the fitted regression \( \exp(\hat{\mu}_i) = \exp(\hat{\beta} \cdot \hat{x}_i) \), and the quantile estimates \( t_i(\alpha) \) and \( t_i(1-\alpha) \).

- Plot of the observed standardized residuals \( \exp(\hat{e}_i) = \exp((\hat{\mu}_i - \hat{\mu}_i)/\hat{\theta}) \) versus the fitted regression.

- A probability plot of the Kaplan and Meier estimate of the residuals \( \hat{e}_i \) with an overlay of the assumed standardized cdf.

b) Plots of the direction of maximum normal curvature vectors: there is a plot for inferences on the entire set of parameters and a plot for inferences on each quantile specified in the QUADATA data set.

c) Optionally, the user can request the creation of up to three data sets.

- A data set containing all the information used to generate the goodness of fit plots. This is requested by specifying a data set name with the macro variable GOODOUT in the macro call. The data set contains the following information. The original time to failure \( T \), the explanatory variables \( X \), the estimates \( \exp(\hat{\mu}_i) \), the quantile estimates \( t_i(\alpha) \) and \( t_i(1-\alpha) \), and standardized residuals \( \exp(\hat{e}_i) \).

- A data set containing all the direction of maximum curvature vectors that were requested. The data set name is given by the DIROUT macro variable. The data set contains an \( h_{max} \) vector for inferences on \( \hat{\beta} \) and an \( h_{max} \) vector for every quantile specified in QUADATA.

- A data set containing the estimated values and the original specification of the quantiles given in QUADATA. To request this data set assign a SAS data set name to the macro variable QUAOUT.

The macro LOCINF requires SAS/IML*, SAS/GRAPH*, and it has been developed for mainframe SAS*.

5 Example

Here we use an example to demonstrate some of the capabilities of the MACRO LOCINF. Our purpose is to illustrate the macro features rather than to describe and interpret the local influence analysis. For a detailed discussion of this example see Escobar and Meeker (1988).

5.1 Data, Models, Objectives

We shall use the "Lung Cancer Survival Data" given in Lawless (1982, p287). These are lung cancer survival times for patients assigned to one of two chemotherapy treatments. Originally the study considered several explanatory variables, but the only one that turned to be significant was "performance status" at diagnosis (a measure of the general medical situation of the patient on a scale of 0 to 100, where large scores are good). Preliminary analysis also suggested that a Weibull or a lognormal model can be used to describe the data (see Prentice 1973 or Lawless 1982).

In this example, we restrict our attention to Weibull and lognormal regression models that include an intercept and the explanatory variable "status". The objectives of the local influence analysis are the identification of those observations that most influence the estimates of

- the entire set of parameters \( (\beta_0, \beta_1, \sigma) \) and

\[ y_{0.5} = \hat{\beta}_0 + \hat{\beta}_1(50) + \sigma e^{-1}(0.5) \] (i.e. the 0.5 quantile of the log(life) distribution for patients with a "status" level equal to 50)

when we introduce a "weight perturbation" into the model. One can also entertain local influence analysis for other inferential objectives such as the estimates of the individual parameters in the model, but we do not pursue them here.
5.2 Macro Call and Output

In Table 1 we show the macro call for the influence analysis with the lognormal model. This call generated Figures 1, 2, and 4-6. Similar calls produced the remaining figures. The independent variable of interest, "status", corresponds to the SAS variable X2 in the input data set and the censoring codes are contained in the FLAG variable. The lognormal analysis was requested through the option DIST=LNORMAL. Some of the macro variables took their default values. In particular, this macro call does not request the data sets needed to regenerate the plots.

Table 1: Macro Call

```sas
DATA A ; INPUT SURTIME FLAG 12-18;
LABEL SURTIME="DAYS' T2='STATUS";
CARDS;
411 0 70 64 8 1 0 0 1
231 0 70 67 18 0 0 0 0
DATA QUANTILS; INPUT X2 PROB;
CARDS;
50 0.6
LOCINF(SURDATA=A,T=SURTIME,X=X2,
CVAR=FLAG,QUADATA=QUANTILS,
P=PROB,DIST=LNORMAL);
```

Figures 1-4 are plots used for assessing the goodness of fit. In the scatter plot, the solid line is the estimated location parameter as a function of the "status" explanatory variable and the dashed lines are the 0.05 and 0.95 quantiles of the distribution as function of "status". Similarly, in Figure 2 the dashed lines are the 0.95, 0.50, and 0.05 quantiles of a standardized lognormal distribution. The lognormal probability plot displays the Kaplan and Meier estimate of the residual and the cdf (the dashed line) of a standardized lognormal distribution. This figure is useful in assessing departures from the lognormal assumption.

Figures 2 and 3 are the same residual plots, but in Figure 3 the individual observations are identified. One can also request the identification of the observations in the Scatter Plot (Figure 1). The identification is requested by setting the macro variable CASEID equal to YES. In Figure 3 some of the ID'S are hard to read because there are several observations with similar residual and predicted values.

The goodness of fit plots do not show any serious model departures and they identify observations #30, #7, #5, and #3 as potential outliers.

Figures 5-6 give index plots of the direction of maximum curvature vectors for inferences on the entire vector of parameters and for the 0.50 lifetime quantile of patients with a "status" equal to 50 when the assumed models are lognormal and Weibull, respectively. Large values of the eigenvector components indicate observations that can have a great deal of influence on the inferences when the observations are not equally weighted. In particular, for the lognormal model and case weight perturbations, observations #7, #5, and #30 have the most potential influence for inferences on the parameter vector.

6 Availability

The SAS macro LOCINF is available from the authors upon request at the address:

Luis A. Escobar
Dept. of Experimental Statistics
Louisiana State University
Baton Rouge, LA 70803-5606

Acknowledgements

We wish to thank Professor Edgar B. Moser for a great deal of suggestions and assistance in writing this macro.

*SAS, SAS/GRAPH and SAS/IML are registered trademarks of SAS Institute, Inc., Cary, NC, USA.

References


Lung Cancer Data
Goodness of Fit Plots—Lognormal Model

Figure 1: Scatter Plot and Regression Line

Figure 2: Residuals Versus Fitted Values

Figure 3: Residuals Versus Fitted Values — Case Identification

Figure 4: Lognormal Probability Plot of Residuals and Lognormal (0,1)
Lung Cancer Data
Directions of Maximum Curvature
Lognormal Model and Case Weight Perturbation

Figure 5: $H_{\text{max}}$ for Full Parameter Vector

Figure 6: $H_{\text{max}}$ for the 0.5 Quantile at Status=50

Lung Cancer Data
Directions of Maximum Curvature
Weibull Model and Case Weight Perturbation

Figure 7: $H_{\text{max}}$ for Full Parameter Vector

Figure 8: $H_{\text{max}}$ for the 0.5 Quantile at Status=50