World Academy of Science, Engineering and Technology
International Journal of Mathematical and Computational Sciences
Vol:3, No:5, 2009

A Renovated Cook's Distance Based On The Buckley-James Estimate In Censored Regression

Nazrina Aziz and Dong Q. Wang

Open Science Index, Mathematical and Computational Sciences Vol:3, No:5, 2009 publications.waset.org/10044/pdf

Abstract—There have been various methods created based on the regression ideas to resolve the problem of data set containing censored observations, i.e. the Buckley-James method, Miller's method, Cox method, and Koul-Susarla-Van Ryzin estimators. Even though comparison studies show the Buckley-James method performs better than some other methods, it is still rarely used by researchers mainly because of the limited diagnostics analysis developed for the Buckley-James method thus far. Therefore, a diagnostic tool for the Buckley-James method is proposed in this paper. It is called the renovated Cook's Distance, (RD_i^*) and has been developed based on the Cook's idea. The renovated Cook's Distance (RD_i^*) has advantages (depending on the analyst demand) over (i) *the change in the fitted value for a single case*, DFIT_i^* as it measures the influence of case i on all n fitted values $\hat{Y}_i^*$ (not just the fitted value for case i as DFIT_i^*) (ii) *the change in the estimate of the coefficient when the ith case is deleted*, DBETA_i^* since DBETA_i^* corresponds to the number of variables p so it is usually easier to look at a diagnostic measure such as RD_i^* since information from p variables can be considered simultaneously. Finally, an example using Stanford Heart Transplant data is provided to illustrate the proposed diagnostic tool.

Keywords—Buckley-James estimators, censored regression, censored data, diagnostic analysis, product-limit estimator, renovated Cook's Distance.

I. Introduction

There have been various methods created based on regression ideas to resolve the problem of data set containing censored observations, i.e. the Buckley-James method, Miller's method, Cox method and Koul-Susarla-Van Ryzin estimators. Miller and Halpern [19] compared the performance of these three methods and found that only the Buckley-James regression method produced reliable estimators for use with censored observations.

In another study, [10] compared several methods of developing estimators in linear regression for a data set with censored observations. The finding is in agreement with [19] whereby the Buckley-James method was selected over the other methods. However, in 1992, they re-examined the Buckley-James and the Cox (the proportional hazards model) methods. The researchers found that the choice of a method relied on the censoring proportion, the form of the failure distribution, the strength of the regression and the form of the censoring distribution.

Later in 2000, [26] described three reasons to support the Buckley-James regression method instead of the Cox method: (i) Most researchers always failed to notice the basic assumptions of Cox method, i.e. normally the assumption was not fulfilled (it might be due to no alternative method in the software resulting in the researcher omitting it); (ii) The Buckley-James method could provide prediction directly from estimators as opposed to the Cox method; (iii) The linear fits resulted from the Buckley-James method are easier to explain to the non-statisticians.

Other than doing comparisons for censored regression estimators that were developed from various methods as to evaluate the performance of Buckley-James method, the diagnostic analysis for Buckley-James method has also attracted a number of researchers as evident in the previous studies. For example, [25] proposed renovated leverage value and renovated scatterplot for censored regression. Later in the year 1999, [24] suggested renovated added variable plot. And finally, renovated partial residual plot, created by Wang, Smith and Aziz [28].

Even though comparison studies show the Buckley-James method performs better than some other methods (see, [10], [19], [26]), it is infrequently used by researchers primarily because of the limited diagnostics analysis developed for the Buckley-James method thus far. Therefore, the current study is designed to develop a diagnostic tool for the Buckley-James method. The proposed diagnostic tool is called the renovated Cook's Distance, (RD_i^*), which is developed based on the Cook's idea.

The Cook's statistics [5] is the best summary of influence due to its tendency to amplify the influence of a case. Therefore, it is chosen to be modified in an attempt to produce the quickest way in detecting the influential case in censored regression, particularly in the Buckley-James method. The renovated Cook's Distance, RD_i^* has advantages (depending on the analyst demand) over

1) $\text{DFIT}_i^* = x_i^T\hat{\beta}^* - x_i^T\hat{\beta}^*{}_{(i)}$ as it measures the influence of case i on all n fitted values $\hat{Y}_i^*$ (not just the fitted value for case i as DFIT_i^*)
2) $\text{DBETA}_i^* = \hat{\beta}^* - \hat{\beta}^*{}_{(i)}$ since DBETA_i^* corresponds to the number of variables, p so it is usually easier to look at a diagnostic measure such as RD_i^* since information from p variables can be considered simultaneously.

DFIT_i^* measures effect of change in fit and DBETA_i^* evaluates change in the estimated regression coefficients for

This research as a part of PhD study supported by the Ministry of Higher Education, Malaysia and Faculty Strategic Grant Research at Victoria University of Wellington.

Nazrina Aziz is a Ph.D. student in School of Mathematics, Statistics and Operations Research, Victoria University of Wellington, New Zealand (e-mail: nazrina.aziz@msor.vuw.ac.nz or nazrina@uum.edu.my).

Dr. Dong Q. Wang is an Associate Professor in School of Mathematics, Statistics and Operations Research, Victoria University of Wellington, New Zealand (e-mail: Dong.Wang@msor.vuw.ac.nz).

Open Science Index, Mathematical and Computational Sciences Vol:3, No:5, 2009 publications.waset.org/10044/pdf

censored regression if the ith row of $X_{n\times(p+1)}$ is deleted.

$\hat{\beta}^*$ represents the coefficients estimated for censored regression of all cases and $\hat{\beta^*_{(i)}}$ is the coefficients estimated for censored regression when the ith row is deleted. The subscript i in parentheses is read as "with case i is removed from $X_{n\times(p+1)}$".

The paper is organized as follows: Section II provides a general idea of the Buckley-James regression and estimation. Section III describes previous diagnostics for Buckley-James censored regression. Section IV explains the proposed diagnostic, renovated Cook's Distance. Finally, Section V provides illustrative examples before presenting the conclusion.

II. Buckley-James regression and estimation

The Buckley-James regession method was proposed by Buckley and James [3]. They modified standard linear regression equations, $y_i = \alpha + \beta x_i + \epsilon_i$, to make it flexible with the data set that possesses censored observations. Let the ith observation have a related censoring time, t_i. Now observed Z_i, δ_i and x_i for $i = 1, 2, \ldots, n$ where

$$Z_i = \min(y_i, t_i)$$

and

$$\delta_i = \begin{cases} 0 \quad \text{(censored)} & \text{if} \quad y_i \geq t_i, \\ 1 \quad \text{(uncensored)} & \text{if} \quad y_i < t_i. \end{cases}$$

Choose the survival time as t_i; if the observation is censored, $\delta_i = 0$ whereas if the observation is uncensored, $\delta_i = 1$, then let the survival time be as y_i. In this method, the old response variable (survival time) needs to be renovated based on their censored status, δ_i.

$$y_i^*(b) = \begin{cases} bx_i + \left[\epsilon_i(b)\delta_i + \hat{E}_b(\epsilon_i(b)|\epsilon_i(b) > c_i(b))(1-\delta_i)\right] & \text{if} \quad \delta_i = 0, \\ y_i & \text{if} \quad \delta_i = 1. \end{cases}$$

The residual is represented by the different types of notation which are c_i and ϵ_i [22].

Let $c_i(b) = t_i - bx_i$ and $\epsilon_i(b) = y_i - bx_i$ and choose $e_i(b) = Z_i - bx_i = \min\{c_i(b), \epsilon_i(b)\}$. Note that

$$\begin{aligned} \hat{E}_b(\epsilon_i(b)|\epsilon_i(b) > c_i(b)) &= \frac{\int_{e_i}^{\infty} \epsilon d\hat{F}_b(\epsilon)}{\int_{e_i}^{\infty} d\hat{F}_b(\epsilon)} \\ &= \sum_{k=1}^{n} w_{ik}(b) e_k(b) \end{aligned} \tag{1}$$

and $w_{ik}(b)$ are the weights developed from the probability mass assigned by F to $e_k(b)$ and Kaplan-Meier estimator $F = 1 - S$ applied to the $e_k(b)$.

Now consider the multivariate censored regression,

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon, \quad \varepsilon \sim F$$

where

- $\mathbf{Y}$ is a $n \times 1$ vector of response variable, which is right censored;
- $\mathbf{X}$ is a known $n \times (p+1)$ matrix as the first column of 1's to provide an intercept;
- β is a $(p+1)\times 1$ vector of parameters where it is estimated by $\mathbf{b^T} = (b_0, b_1, \ldots, b_p)$;
- ε is $n \times 1$ vector of errors and the distribution has an unknown survival function, $S = 1 - F$.

First the renovated response variable needs to be obtained as the linear censored regression. This can be done using the following equation

$$\mathbf{Y}^*(\mathbf{b}) = \mathbf{Xb} + \mathbf{Q}(\mathbf{b})(\mathbf{Z} - \mathbf{Xb}). \tag{2}$$

Next, the Buckley-James estimators can be developed as follows

$$\hat{\beta} = (\mathbf{X^T W X})^{-1}\mathbf{X^T W Y}^* \tag{3}$$

where W is the upper triangle Renovation Weight Matrix [22] containing censored status on the main diagonal as below

$$\begin{aligned} W(\mathbf{b}) &= \text{diag}(\delta) + \{w_{ik}(\mathbf{b})\} \\ &= \begin{pmatrix} \delta_1 & w_{12}(\mathbf{b}) & w_{13}(\mathbf{b}) & \ldots & w_{1n}(\mathbf{b}) \\ 0 & \delta_2 & w_{23}(\mathbf{b}) & \ldots & w_{2n}(\mathbf{b}) \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \ddots & w_{(n-1)n}(\mathbf{b}) \\ 0 & 0 & 0 & \ldots & \delta_n \end{pmatrix} \end{aligned} \tag{4}$$

where

$$w_{ik}(\mathbf{b}) = \begin{cases} \dfrac{d\hat{F}(e_k(\mathbf{b}))\delta_k(1-\delta_i)}{\hat{S}(e_i(\mathbf{b}))} & \text{if} \quad k > i, \\ 0 & \text{if} \quad \text{otherwise.} \end{cases} \tag{5}$$

In fact, various efforts have been carried out as illustrated in the previous studies to improve the Buckley-James method (see, [15], [16], [18]).

In addition to Buckley-James estimators, previous studies also mentioned the various diagnostic tools on censored regression (see, [21], [24], [25], [28]). This is further described in the following section.

III. Diagnostic analysis for Buckley-James censored regression

There are various techniques to examine a model and discover the outlying and influential observations in regression with a common data set (details can be found in [2], [4], [6]). Thus, in censored regression, particularly the estimators estimated using the Buckley-James method, a few diagnostic tools can also be found.

A. Renovated Scatterplot

The new response variable (Y^*), particularly for censored observations can be obtained after finding the solution to the Buckley-James estimator. Note that the response variable for uncensored observations would remain the same. By using Y^*, now the scatterplot of X vs Y^* can be developed. This means, the plot contains renovated points and uncensored points.

Next, [14] made an effort to develop residual plot similar to the standard residual plot for standard regression. This plot was developed by using modified residuals to examine heteroscedacity and the violation of the other distributional assumptions. The modified residual is given by

$$e_i^* = \delta_i(Y_i - x_i^T\hat{\beta}) + (1-\delta_i)D_i, \tag{6}$$

where D_i is randomly generated from the conditional distribution estimated from the fitted model [9].

Open Science Index, Mathematical and Computational Sciences Vol:3, No:5, 2009 publications.waset.org/10044/pdf

B. Renovated Added Variable Plot

The added variable plots are the diagnostic tools that permit evaluation of the role of individual variables within the multiple regression model. They are used to visually assess (i) whether a variable should be included in the model, (ii) the presence of outliers and influential cases, and (iii) the possibility of non-linear relationship between Y and individual X in the model. An added variable plot is a way to look at the marginal role of variable X_p in the model, given that other independents are already in the model.

Smith and Peiris [24] proposed the renovated added variable plot for censored regression. Assume the censored regression model, $Y^* = \beta_0 + \beta_1 X_1 + \beta_2 X_2$, the renovated added variable plot for censored regression can be defined in terms of residuals as the plot $e^*(Y^*|X_1)$ against $e^*(X_2|X_1)$, where $e^*(Y^*|X_1)$ is the renovated residual (Y regress on X_1) and $e^*(X_2|X_1)$ is the renovated residual (X_2 regress on X_1).

It can be shown that the slope of the added variable plot of $e^*(Y^*|X_1)$ on $e^*(X_2|X_1)$ is equal to the estimated coefficient β_2 of X_2 in the censored regression model $Y^* = \beta_0 + \beta_1 X_1 + \beta_2 X_2$ (see, [24]).

C. Renovated Partial Residual Plot

Partial residual plots examine whether the linearity assumption in a multiple regression model appears to be satisfied. Let $Y^* = \beta_0 + \beta_1 X_1 + \beta_2 X_2$ be the censored regression model, [28] defined the renovated partial residual vector for X_2 as $R^*_{X_2} = (I - H^*)Y^* + X_2\beta_2$, where H^* is the renovated hat matrix. From the plot, if the point lies very close to the straight lines, it suggests the X_2 affects the Y^* strength linearity. Wang, Smith and Aziz [28] also proved that the slope of the renovated partial residual plot is equal to the β_2 in Y^*.

D. Renovated Hat Matrix

The hat matrix is used to identify the outlying observations. In 1995, [25] proposed the renovated hat matrix, H^* for censored regression. H^* is developed from Lemma 2.1 in [4]. The renovated hat matrix for censored regression is given by

$$H^* = X(X^T W X)^{-1} X^T W.$$

Next, the vector of renovate residual can be defined as $e^* = Y^* - \hat{Y}^* = Y^* - H^* Y^*$, so that $e^* = (I - H^*)Y^*$. The H^* is not symmetric, however it fulfils $(H^*)^2 = H^*$, $(I - H^*)^2 = I - H^*$, $\mathrm{tr}(H^*) = p$ and $H^*(Y^* - X\beta) = 0$. It follows that the variance of the renovated residual estimate is $\sigma^2(e^*) = \sigma^2(I - H^*)$. Thus, the variance of an individual renovated residual, e_i^*, is $\sigma^2(e_i^*) = \sigma^2(1 - h_{ii}^*)$ and h_{ii}^* can be calculated without calculating the whole H^*,

$$h_{ii}^* = x_i^T (X^T W X)^{-1} X^T w_i,$$

where w_i is the $n \times 1$ vector of the weights estimated in (5). h_{ii}^* measures the leverage of an observation. The high-leverage observation can be identified by comparing the h_{ii}^* value with $2p/n$. In censored regression, the h_{ii}^* is equal to zero for $\delta_i = 0$, i.e. censored observation. In case of $\delta_i = 1$, if the $h_{ii}^* > 2p/n$, then the observation could be flagged as uncommonly large.

IV. Renovated Cook's Distance for Buckley-James Censored Regression

A case becomes influential if, when it is excluded from the regression, causes a substantial change in the estimated regression function. This can be measured by calculating DFIT_i^*.

DFIT_i^* for Buckley-James model measures the influence of case i on its own fitted value, $\hat{Y}_i^*$. DFIT_i^* is given by Smith [22] as

$$\mathrm{DFIT}_i^* = \frac{h_{ii}^* \epsilon_i^*}{(1 - h_{ii}^*)} \tag{7}$$

where $\epsilon_i^* = Y_i^* - x_i^T \hat{\beta}^*$ and $h_{ii}^* = x_i^T (X^T W X)^{-1} X^T w_i$.

DFIT_i^* represents the number of estimated standard deviations of $\hat{Y}_i^*$ where the fitted value $\hat{Y}_i^*$ increases or decreases with the inclusion case i in regression.

As indicated earlier, this paper aims to propose the renovated Cook's Distance, RD_i^*. It measures the influence of case i on all n fitted values $\hat{Y}_i^*$ (not just the fitted values for case i as $DFIT_i^*$). In a general version of Cook's Distance for least square regression (LSR), one can have

$$D_i = \frac{(\hat{Y} - \hat{Y}_{(i)})^T (\hat{Y} - \hat{Y}_{(i)})}{p\sigma^2} \tag{8}$$

where $\hat{Y}_{(i)}$ is the deleted fitted value when the ith point is deleted. To produce a renovated Cook's Distance for censored regression, let the Buckley-James estimators be

$$\hat{\beta}^* = (X^T W X)^{-1} (X^T W Y^*).$$

Therefore, the Buckley-James estimators without ith observation is given as,

$$\begin{aligned}
\hat{\beta}^*_{(i)} &= (X_{(i)}^T W_{(i,i)} X_{(i)})^{-1} (X_{(i)}^T W_{(i,i)} Y_{(i)}^*) \\
&= (X_{(i)}^T W_{(i,i)} X_{(i)})^{-1} (X^T W Y^* - x_i w_i^T Y^*) \\
&= \left\{ (X^T W X)^{-1} + \frac{(X^T W X)^{-1} X^T w_i x_i^T (X^T W X)^{-1}}{1 - x_i^T (X^T W X)^{-1} X^T w_i} \right\} \\
&\qquad \left\{ X^T W Y^* - x_i w_i^T Y^* \right\} \\
&= \left\{ (X^T W X)^{-1} X^T W Y^* + \frac{(X^T W X)^{-1} X^T w_i x_i^T (X^T W X)^{-1} X^T W Y^*}{1 - x_i^T (X^T W X)^{-1} X^T w_i} \right\} \\
&\qquad - \left[\left\{ (X^T W X)^{-1} + \frac{(X^T W X)^{-1} X^T w_i x_i^T (X^T W X)^{-1}}{1 - x_i^T (X^T W X)^{-1} X^T w_i} \right\} X^T w_i y_i^* \right] \\
&= \left\{ \hat{\beta}^* + \frac{(X^T W X)^{-1} X^T w_i x_i^T \hat{\beta}^*}{1 - x_i^T (X^T W X)^{-1} X^T w_i} \right\} - \left\{ \frac{(X^T W X)^{-1} X^T w_i y_i^*}{1 - x_i^T (X^T W X)^{-1} X^T w_i} \right\}
\end{aligned} \tag{9}$$

where $X_{(i)}$ and $Y_{(i)}^*$ denote the $\mathbf{X}$ matrix and the response variable respectively when the ith row removed and

$$W_{(i,i)} = \begin{pmatrix}
\delta_1 & w_{12} & w_{13} & \cdots & w_{1(i-1)} & w_{1(i+1)} & \cdots & w_{1n} \\
 & \delta_2 & w_{23} & \cdots & w_{2(i-1)} & w_{2(i+1)} & \cdots & w_{2n} \\
 & & \ddots & \ddots & \vdots & \vdots & \vdots & \vdots \\
 & & & \ddots & \ddots & & & \\
 & & & & \delta_{(i-1)} & w_{(i-1)(i+1)} & \cdots & w_{(i-1)n} \\
 & & & & & \delta_{(i+1)} & \cdots & w_{(i+1)n} \\
 & & & & & & \ddots & \vdots \\
0 & & & & & & & \delta_n
\end{pmatrix}$$

is the upper triangle Renovation Weight Matrix when ith row and column are deleted from the matrix. Let

$$h_{ii}^* = x_i^T (X^T W X)^{-1} X^T w_i$$

Open Science Index, Mathematical and Computational Sciences Vol:3, No:5, 2009 publications.waset.org/10044/pdf

be a renovated leverage for censored regression, hence replaces h_{ii}^* in (9), so the Buckley-James estimators without ith observation can be defined as

$$\begin{aligned}\hat{\beta}_{(i)}^* &= \left\{\hat{\beta}^* + \frac{(X^TWX)^{-1}X^Tw_ix_i^T\hat{\beta}^*}{1-h_{ii}^*}\right\} - \left\{\frac{(X^TWX)^{-1}X^Tw_iy_i^*}{1-h_{ii}^*}\right\}\\ &= \hat{\beta}^* + \left\{(X^TWX)^{-1}X^Tw_i\right\}\left\{\frac{(x_i^T\hat{\beta}^*)-y_i^*}{1-h_{ii}^*}\right\}\\ &= \hat{\beta}^* - \left\{(X^TWX)^{-1}X^Tw_i\right\}\left\{\frac{y_i^*-\hat{y}_i^*}{1-h_{ii}^*}\right\}\\ &= \hat{\beta}^* - \left\{(X^TWX)^{-1}X^Tw_i\right\}\left\{\frac{\hat{e}_i^*}{1-h_{ii}^*}\right\}.\end{aligned}$$

Now the renovated Cook's Distance for censored regression can be developed as

$$\begin{aligned}RD_i^* &= \frac{(\hat{\beta}_{(i)}^*-\hat{\beta}^*)^TS^*(\hat{\beta}_{(i)}^*-\hat{\beta}^*)}{ps^2}\\ &= \frac{(\hat{e}_i^*)^2}{ps^2}\left\{\frac{w_i^TX(X^TWX)^{-1}X^Tw_i}{(1-h_{ii}^*)^2}\right\}\\ &= \frac{(\hat{e}_i^*)^2}{ps^2}\left\{\frac{h_{ii}^{**}}{(1-h_{ii}^*)^2}\right\},\end{aligned}$$

where $S^* = X^TWX$, s^2 is estimate variance and $h_{ii}^{**} = w_i^TX(X^TWX)^{-1}X^Tw_i$ and $\hat{e}_i^* = y_i^* - \hat{y}_i^*$.

***Theorem** 1:* The renovated leverage of an observation in censored regression, h_{ii}^*, can be presented in the following form, $w_i^TX(X^TWX)^{-1}X^Tw_i$, which is defined as h_{ii}^{**}. Therefore, $h_{ii}^{**} = h_{ii}^*$.

Proof: Let the renovated leverage be

$$H^* = X(X^TWX)^{-1}X^TW$$

and $X = (X_1 \;\; X_2)$ where X_1 is an $(n \times r)$ matrix of rank r and X_2 is an $n \times (k-r)$ matrix of rank $k-r$. From [24], one can find

$$H^* = H_1^* + (I - H_1^*)(X_2MX_2^TW)(I - H_1^*)$$

where $H_1^* = X_1(X_1^TWX_1)^{-1}X_1^TW$ and

$$M = [X_2^TW(I - H_1^*)X_2]^{-1}.$$

By using Lemma 2.1 in [4], H^{**} can be developed as below

$$\begin{aligned}H^{**} &= WX(X^TWX)^{-1}X^TW\\ &= \begin{pmatrix}WX_1 : & WX_2\end{pmatrix}\begin{pmatrix}X_1^TWX_1 & X_1^TWX_2\\ X_2^TWX_1 & X_2^TWX_2\end{pmatrix}^{-1}\begin{pmatrix}X_1^TW\\ X_2^TW\end{pmatrix}\\ &= \begin{pmatrix}WX_1 : & WX_2\end{pmatrix}\begin{pmatrix}v_{11} & v_{12}\\ v_{21} & M\end{pmatrix}\begin{pmatrix}X_1^TW\\ X_2^TW\end{pmatrix}\\ &= WX_1(X_1^TWX_1)^{-1}X_1^TW + (I - H_1^*)(X_2MX_2^TWW)(I - H_1^*)\\ &= X_1(X_1^TWX_1)^{-1}X_1^TWW + (I - H_1^*)(X_2MX_2^TWW)(I - H_1^*)\\ &= X_1(X_1^TWX_1)^{-1}X_1^TW^2 + (I - H_1^*)(X_2MX_2^TW^2)(I - H_1^*),\end{aligned}$$

where

$$\begin{aligned}v_{11} &= (X_1^TWX_1)^{-1} + (X_1^TWX_1)^{-1}(X_1^TWX_2)M(X_2^TWX_1)(X_1^TWX_1)^{-1};\\ v_{12} &= -(X_1^TWX_1)^{-1}(X_1^TWX_2)M;\\ v_{21} &= -M(X_2^TWX_1)(X_1^TWX_1)^{-1}.\end{aligned}$$

From the properties of the weight matrix, it is known that $W^2 = W$, idempotence, see the proof in [23]. Hence,

$$\begin{aligned}H^{**} &= X_1(X_1^TWX_1)^{-1}X_1^TW^2 + (I - H_1^*)(X_2MX_2^TW^2)(I - H_1^*)\\ &= X_1(X_1^TWX_1)^{-1}X_1^TW + (I - H_1^*)(X_2MX_2^TW)(I - H_1^*)\\ &= H_1^* + (I - H_1^*)(X_2MX_2^TW)(I - H_1^*)\\ &= H^*\end{aligned}$$

since the renovated leverage, h_{ii}^*, comprises the diagonal entries of H^*, therefore $h_{ii}^{**} = h_{ii}^*$. ■

Based on Theorem 1, Theorem 2 is given as follows:

***Theorem** 2:* The renovated Cook's Distance is given by

$$RD_i^* = \frac{(\hat{e}_i^*)^2}{ps^2}\left\{\frac{h_{ii}^*}{(1-h_{ii}^*)^2}\right\},$$

where $h_{ii}^* = x_i^T(X^TWX)^{-1}X^Tw_i$.

The formula shows that RD_i^* is large when either renovated residual, e_i^*, or the renovated leverage, h_{ii}^*, is large, or both. It should be noted that due to censoring estimates of the residual variance, s^2 could easily inflate the RD_i^*. This problem is solved by calculating s^2 using the variance estimator proposed by Smith [20]. Simulation studies by [12] and [13] showed that Smith estimator performed the best. The variance estimator by Smith [20] is given by

$$\hat{\sigma}_{SMITH}^2 = \frac{n_u}{n_u-2}g^{-2}\Big[\sum_{i=1}^{n}(x_i-\bar{x})^2\tilde{\sigma}_i^2\Big] \qquad (10)$$

where n_u is the number of uncensored observations, $\tilde{\sigma}_i^2$ and g are defined by

$$\tilde{\sigma}_i^2 = \int \epsilon^2 d\hat{F}_{\hat{\beta}}(\epsilon) - (1-\delta_i)\left[\frac{\int_{e_i}^{\infty}\epsilon^2 d\hat{F}_{\hat{\beta}}(\epsilon)}{\int_{e_i}^{\infty}d\hat{F}_{\hat{\beta}}(\epsilon)} - \left\{\frac{\int_{e_i}^{\infty}\epsilon d\hat{F}_{\hat{\beta}}(\epsilon)}{\int_{e_i}^{\infty}d\hat{F}_{\hat{\beta}}(\epsilon)}\right\}^2\right]$$

where

$$\int \epsilon^2 d\hat{F}_{\hat{\beta}}(\epsilon) = \frac{1}{n}\sum_{i=1}^{n}\left[\delta_i[e_i(b)]^2 + (1-\delta_i)\sum_{k=1}^{n}w_{ik}(b)[e_k(b)]^2\right]$$

and

$$g = \sum_{i=1}^{n}(x_i-\bar{x})^2\Big[1-(1-\delta_i)\hat{p}_i(b)\Big]$$

where

$$\hat{p}_i(b) = 1 + \hat{\lambda}(e_i)\left[e_i - \frac{\int_{e_i}^{\infty}\epsilon d\hat{F}_{\hat{\beta}}(\epsilon)}{\int_{e_i}^{\infty}d\hat{F}_{\hat{\beta}}(\epsilon)}\right]$$

$\hat{\lambda}(e_i)$ is estimated hazard function for e_i and it is calculated using the life table method as in [17]. Influence cases can easily be detected by using the index plot $\{i, RD_i^*\}$ where i is the case number; particularly influential observations that belong to the uncensored group.

Recall that the original Cook's Distance will flag observations in standard regression from a normal data i.e. uncensored data that is greater than 1 or 2 as influential (see, [27]). The principle occurs due to the argument that the Cook's Distance does not have an F-distribution (see, [4]). As such, the uncensored data using the renovated Cook's Distance will also be given extra attention if their RD_i^* value is greater than 1 or 2.

In censored regression, it is noted that the RD_i^* is equal to zero for observation with $\delta_i = 0$ i.e. censored observation. This follows from h_{ii}^*, recall that $h_{ii}^* = 0$ for censored observations. Even though the circumstances agree well with Weissfeld and Schneider [29] as censored observations have

Open Science Index, Mathematical and Computational Sciences Vol:3, No:5, 2009 publications.waset.org/10044/pdf

a high tendency to be less influential than uncensored observations, one still has to be aware of the potency of censored observations to influence the censored regression. Aziz and Wang [1] discusses and presents a new diagnostic tool based on local influence to overcome this issue.

V. Result

The Stanford Heart Transplant data set is a standard data set for censored regression. It is taken from a Stanford Heart Transplant program which began in October 1967. It has had a number of versions since then. In this paper, the data is taken from R library, but data on patients who were admitted to the program but did not receive the transplant have been omitted.

Therefore, for regression and diagnostic analysis, 69 patients were used and of these 69 patients, 45 deceased and hence were uncensored while 24 are still alive, and hence were censored. The explanatory variables were age in years and censored status. Since the data for the age was given in days, it was divided by 365. A patient who died on the same day during his/her transplant was given a survival time of one day. The response variable was time survival; this variable has been transformed to log base 10, as the linear model is often appropriate when the response variable is measured on the logarithm scale [3]. Details about this data set can be found in [7].

The appendix shows a value of $\hat{e}_i, h_{ii}, h^*_{ii}, RD^*_i$ for each observation. The values of h_{ii} and h^*_{ii} are in agreement with Weissfeld and Schneider [25]. The h^*_{ii} value is equal to zero for all patients with censored data $(\delta_i = 0)$. Next, the RD^*_i value of each observation in Appendix A was scrutinized . The youngest uncensored patients of 19.6 years (case 17), did not give the largest value of RD^*_i even though this observation showed the highest value of h^*_{ii}. Case 5, which is the uncensored patient of age 29.2, gave the highest value of RD^*_i. This patient had a higher residual value than the youngest patient.

The plot of renovated leverage in Fig. 1 clearly represents the youngest patient of 19.6 years and patient of the age 29.2 years as the two cases with the largest h^*_{ii}. Now refer to Fig. 2, which is the plot of the renovated Cook's Distance. The figure shows similar cases showing the two largest values of RD^*_i, with the patient of age 29.2 years leading. Other patients with RD^*_i value larger than 1 are from case 2, 3, 4 and 69 corresponding to patients aged 41.5, 54.1, 40.3 and 54.0 years.

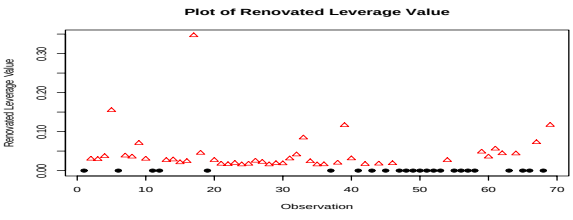


Fig. 1. Renovated leverage plot for Stanford Heart Transplant data where the triangle represents uncensored observation and the circle represents censored observation

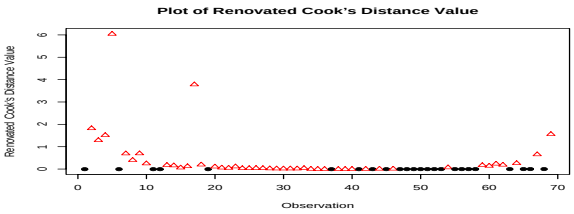


Fig. 2. Renovated Cook's Distance plot for stanford heart transplant data where the triangle represents uncensored observation and the circle represents censored observation

VI. Conclusion

The result of the modified Cook's statistics, without doubt, clearly shows influence cases for the censored regression. Note that the censored points cannot be influential cases as the points have no renovated leverage $(h^*_{ii} = 0)$, it follows that RD^*_i is also equal to zero. This issue needs further inspection due to the concern of the possibility of censored points to become influential case in censored regression. In [1], a new diagnostic tool will be proposed to solve this problem using the local influence approach.

As opposed to censored points, uncensored points tend to be more influential when the RD^*_i is used as a diagnostic tool. However, it is noted that one cannot simply consider the point with the highest leverage as the most influential case. Recall $e^*_i = (1 - h^*_{ii})y^*_i$, obviously, the larger h^*_{ii}, the smaller e^*_i and it follows that the value of RD^*_i will decrease. From the result, the Stanford Heart Transplant data example shows that the youngest patient with the highest leverage value did not emerge as the most influential case in this data set. The proposed diagnostics tool, RD^*_i can be considered as the easiest way to detect the influential cases in censored regression and produce a comparable result with Smith and Peiris [24] and Weissfeld and Schneider [25].

Appendix A
Details Information of the Stanford Heart Transplant Data Based on Age, $\hat{e}_i, \delta_i, h_{ii}, h^*_{ii}, h^{**}_{ii}, RD^*_i$

Cases	Age	$\hat{e}_i$	δ_i	h_{ii}	$h^*_{ii} = h^{**}_{ii}$	RD^*_i
1	35.1	-2.620	0	0.027	0.000	0.000
2	41.5	-2.440	1	0.021	0.030	1.881
3	54.1	-2.086	1	0.019	0.029	1.334
4	40.3	-1.996	1	0.017	0.037	1.565
5	29.2	-1.706	1	0.015	0.155	6.239
6	28.6	-1.688	0	0.028	0.000	0.000
7	40.3	-1.327	1	0.026	0.038	0.718
8	55.3	-1.052	1	0.017	0.035	0.417
9	36.2	-0.945	1	0.036	0.071	0.722
10	54.3	-0.904	1	0.030	0.030	0.257
11	23.6	-0.901	0	0.016	0.000	0.000
12	45.0	-0.864	0	0.016	0.000	0.000
13	42.8	-0.812	1	0.043	0.027	0.184
14	42.5	-0.749	1	0.021	0.028	0.165
15	52.1	0.556	1	0.045	0.021	0.066
16	53.0	-0.719	1	0.027	0.024	0.128
17	19.6	-0.697	1	0.015	0.347	3.903

Open Science Index, Mathematical and Computational Sciences Vol:3, No:5, 2009 publications.waset.org/10044/pdf

continue

Cases	Age	$\hat{e}_i$	δ_i	h_{ii}	$h_{ii}^* = h_{ii}^{**}$	RD_i^*
18	56.9	-0.645	1	0.078	0.045	0.204
19	26.7	-0.083	0	0.016	0.000	0.000
20	53.8	-0.632	1	0.020	0.027	0.112
21	46.3	-0.606	1	0.016	0.017	0.064
22	47.1	-0.575	1	0.059	0.016	0.054
23	45.3	0.769	1	0.018	0.019	0.117
24	49.0	-0.497	1	0.018	0.015	0.039
25	50.6	-0.476	1	0.016	0.017	0.040
26	53.3	-0.427	1	0.015	0.025	0.047
27	52.5	-0.423	1	0.033	0.022	0.041
28	49.1	-0.413	1	0.016	0.015	0.027
29	51.3	-0.345	1	0.015	0.019	0.023
30	51.1	-0.337	1	0.031	0.018	0.021
31	54.6	-0.265	1	0.014	0.031	0.023
32	56.4	-0.222	1	0.016	0.041	0.022
33	61.5	-0.206	1	0.015	0.084	0.042
34	43.9	-0.166	1	0.022	0.024	0.007
35	48.0	-0.153	1	0.024	0.015	0.004
36	47.4	-0.107	1	0.015	0.016	0.002
37	26.7	-0.633	0	0.019	0.000	0.000
38	51.8	-0.017	1	0.016	0.020	0.000
39	64.5	-0.015	1	0.048	0.116	0.000
40	42.7	0.067	1	0.016	0.031	0.001
41	47.8	0.112	0	0.020	0.000	0.000
42	48.8	0.169	1	0.145	0.017	0.005
43	32.7	0.224	0	0.015	0.000	0.000
44	49.5	0.232	1	0.015	0.017	0.010
45	48.7	0.247	0	0.023	0.000	0.000
46	48.0	0.251	1	0.015	0.019	0.012
47	46.5	0.360	0	0.084	0.000	0.000
48	49.0	0.361	0	0.034	0.000	0.000
49	38.8	0.426	0	0.067	0.000	0.000
50	54.4	0.453	0	0.021	0.000	0.000
51	36.7	0.469	0	0.016	0.000	0.000
52	41.4	0.481	0	0.026	0.000	0.000
53	47.4	0.496	0	0.015	0.000	0.000
54	48.8	0.507	1	0.023	0.027	0.071
55	52.9	0.523	0	0.024	0.000	0.000
56	52.1	-0.727	0	0.017	0.000	0.000
57	48.0	0.562	0	0.016	0.000	0.000
58	33.2	0.576	0	0.015	0.000	0.000
59	44.9	0.578	1	0.027	0.048	0.175
60	50.9	0.620	1	0.019	0.035	0.144
61	43.4	0.624	1	0.021	0.056	0.241
62	45.9	0.637	1	0.015	0.044	0.193
63	40.6	0.725	0	0.015	0.000	0.000
64	48.6	0.757	1	0.015	0.043	0.269
65	45.3	-0.520	0	0.021	0.000	0.000
66	48.5	0.893	0	0.084	0.000	0.000
67	58.4	0.899	1	0.109	0.072	0.671
68	48.9	0.955	0	0.071	0.000	0.000
69	54.0	1.042	1	0.037	0.117	1.607

REFERENCES

[1] Aziz, N. and Wang, D. Q. (2009). Local influence for Buckley-James censored regression, submitted for publication to the Communication in Statistics-Theory and Method.

[2] Belsley, D. A., Kuh, E. and Welsch, R. E. (1980). Regression diagnostics identifying influential data and sources of colinearity, John Wiley & Sons, New York.

[3] Buckley, J. and James, I. (1979). Linear regression with censored data, Biometrika 66(3): 429-436.

[4] Chatterjee, S. and Hadi, A. S. (1988). Sensitivity analysis in linear regression, John Wiley, United States.

[5] Cook, R. D. (1977). Detection of influential observation in linear regression, Technometrics 19(1).

[6] Cook, R. D. and Weisberg, S. (1982). Residuals and Influence in regression, Chapman and Hall, New York.

[7] Crowley, J. and Hu, M. (1977). Covariance analysis of heart transplant survival data, Journal of the American Statistical Association 72(357): 27-36.

[8] Currie, I. D. (1996). A note on Buckley-James estimators for censored data, Biometrika 83(4): 912-915.

[9] Glasson, S. (2007). Censored Regression Techniques for Credit Scoring, PhD thesis, RMIT University.

[10] Heller, G. and Simonoff, J. S. (1990). A comparison of estimators for regression with a censored response variable, Biometrika 77(3): 515-520.

[11] Heller, G. and Simonoff, J. S. (1992). Prediction in censored survival data: A comparison of the proportional hazards and linear regression models, Biometrika 48(1): 101-115.

[12] Hillis, S. L. (1993). A comparison of three Buckley-James variance estimators, Communication in Statistics B 22(4): 955-973.

[13] Hillis, S. L. (1994). A heuristic generalisation of smith's Buckley-James variance estimator, Communications in statistics. Simulation and computation 23: 713-831.

[14] Hillis, S. L. (1995). Residual plots for the censored data linear regression model, Statistics in Medicine 14: 2023-2036.

[15] James, I. R. and Smith, P. J. (1984). Consistency results for linear regression with censored data, The Annals of Statistics 12(2): 590-600.

[16] Lai, T. L. and Ying, Z. (1991). Large sample theory of a modified Buckley-James estimator for regression analysis with censored data, The Annals of Statistics 19(3): 1370-1402.

[17] Lee, E. T. (1980). Statistical methods for survival data analysis, Lifetime Learning, California.

[18] Lin, J. S. and Wei, L. J. (1992). Linear regression analysis based on Buckley-James estimating equation, Biometrics 48(3): 679-681.

[19] Miller, R. and Halpern, J. (1982). Regression with censored data, Biometrika 69(3): 521-531.

[20] Smith, P. J. (1986). Estimation in linear regression with censored response, Pacific Statistical Congress, Amsterdam, Holland, pp. 261-265.

[21] Smith, P. J. (1995). On plotting renovated samples, Biometrics 51: 1147-1151.

[22] Smith, P. J. (2002). Analysis of failure and survival data, Chapman & Hall, United States.

[23] Smith, P. J. (2004). Using linear regression techniques with censored data, International Journal of Reliability, Quality and Safety Engineering 11(2): 163-173

[24] Smith, P. J. and Peiris, L. W. (1999). Added variable plots for linear regression with censored data, Communication in Statistics-Theory and Method 28(8): 1987-2000.

[25] Smith, P. J. and Zhang, J. (1995). Renovated scatterplots for censored data, Biometrika 82(2): 447-452.

[26] Stare, J., Heinzl, H. and Harrell, F. (2000). On the use of buckley and james least squares regression for survival data, New Approach in Applied Statistics 12: 125-134.

[27] Velleman, P. F. and Welsch, R. E. (1981). Efficient computing of regression diagnostics, The American Statistician 35(4): 234-242.

[28] Wang, D. Q., Smith, P. J. and Aziz, N. (2009). Renovated partial residuals and properties for censored regression, submitted for publication to the Computational Statistics and Data Analysis.

[29] Weissfeld, L. A. and Schneider, H. (1990). Influence diagnostics for the normal linear model with censored data, Australian Journal Statistics 32(1): 11-20.