Application of Local Influence Diagnostics to the Buckley-James Model

Nazrina Aziz[1] and Dong Q Wang[2]

[1]Universiti Utara Malaysia
[2]Victoria University of Wellington, New Zealand

19th International Conference on Computational Statistics
Paris, France
August 22-28, 2010

Introduction

- Buckley-James Model?

  Method used to resolve the problem of a data set containing censored observations
- Censored observations?

  A data set that contains observations with incomplete information occurs when the event of interest is not observed
- Some examples:
  - In biological research: the time from diagnosis to death
  - For industrial research: an example of special interest can be the life time of machine components.
- Are there any other approaches?

  Methods used are based on regression ideas, i.e the Miller's method, the Cox method and Koul-Susarla-Van Ryzin estimators

Outline

Motivation

- Which methods perform better?

Study	Preference
Miller & Halpern (1982)	The Buckley-James method
Heller & Simonoff (1990)	The Buckley-James method
Heller & Simonoff (1992)	The Buckley-James & Cox method
Stare, Heinzl & Harrell (2000)	The Buckley-James method

- Current study: Buckley-James method
- But it is rarely used. Why?
  - Limited diagnostics analysis developed for the Buckley-James method
  - In the previous diagnostics of Buckley-James model, influential observations merely come from uncensored observations in the data set.

Outline

Objective of the study

- Solution?

  Current study is designed to develop a diagnostic tool for the Buckley-James method

- What is the new diagnostic tool?

  The local influence diagnostics for the Buckley-James model, which consist of
  - variance perturbation;
  - response variable perturbation;
  - independent variables perturbation.

- Advantages

  The proposed diagnostics improves the previous ones by taking into account both censored and uncensored data to have a possibility to become an influential observation

Buckley-James censored regression

- Who introduced Buckley-James method?

  Buckley and James in 1979

- How does it works?
  - Modify the least square standard equations to make it suitable for a data set exposed to censored observations.
  - First, review the standard linear regression with the complete data set:

$$Y_i = \alpha + \beta x_i + \varepsilon_i$$

  - Let the *ith* observation have a related censoring time, t_i.
  - Now observe Z_i, δ_i and x_i for $i = 1, 2, \ldots, n$ where

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

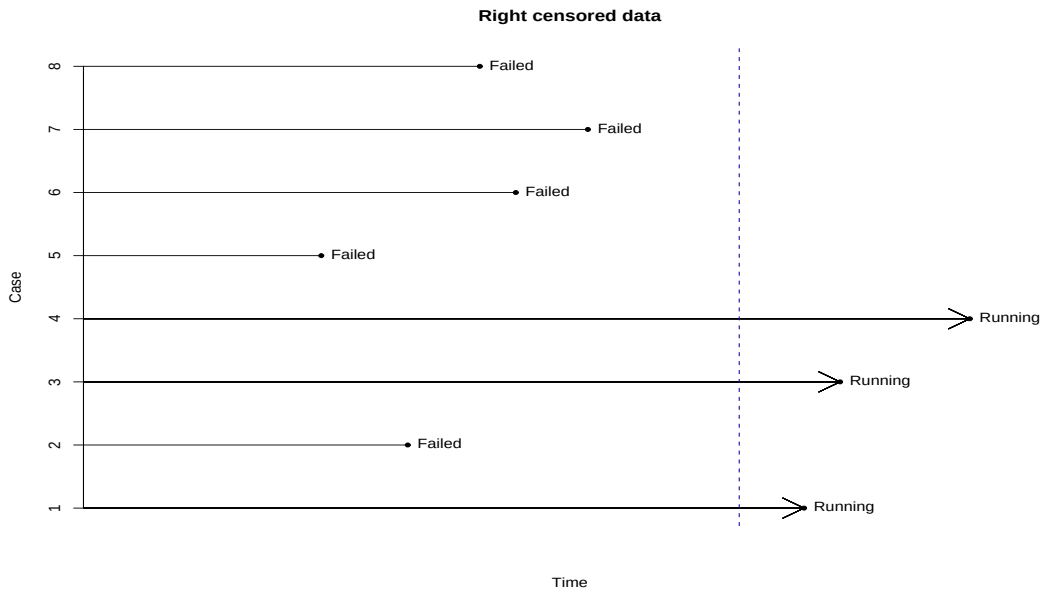

Figure: Plot of right censored data with the dashed lines representing the time line

- and

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

- Renovate the old response variable (survival time) based on its 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 $c_i(b) = t_i - bx_i$ or $\epsilon_i(b) = y_i - bx_i$. Choose $e_i(b) = \min\{c_i(b), \epsilon_i(b)\}$
- Note that

$$\hat{E}_b(\epsilon_i(b)|\epsilon_i(b) > c_i(b)) = \frac{\int_{e_i}^{\infty} \epsilon d\hat{F}_b(\epsilon)}{\int_{e_i}^{\infty} d\hat{F}_b(\epsilon)}$$

$$= \sum_{k=1}^{n} w_{ik}(b) e_k(b)$$

- Next, one can develop the Buckley-James estimator of β as follows

$$\sum_{i=1}^{n}(x_i - \bar{x})(Y_i^* - x_i\hat{\beta}) = 0.$$

- By using the iteration, first get the initial estimate of the slope, $\hat{\beta}^{(0)}$, then the Buckley-James estimator of β can be obtained as below

$$\frac{\sum_{i=1}^{n}(x_i - \bar{x})Y_i^*(\hat{\beta}_n)}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \hat{\beta}_{n+1}$$

  where $\hat{\beta}_n$ is the estimate of β for the *nth* iteration, $n = 1, 2, \ldots$.
- The iteration is stopped when $|\hat{\beta}_{n+1} - \hat{\beta}_n|$ is small and reaches convergence.
- Later one can estimate $\hat{\alpha}$ as follows

$$\hat{\alpha} = \frac{Y^*(\hat{\beta}) - \hat{\beta}x_i}{n} \tag{1}$$

Multivariate censored regression

- What about multivariate censored regression?

  Consider $\mathbf{Y} = \mathbf{X}\beta + \varepsilon, \quad \varepsilon \sim F$
- How does it work?
  - First, the renovated response variable needs to be obtained as the linear censored regression,

$$\mathbf{Y}^*(\mathbf{b}) = \mathbf{Xb} + \mathbf{W}(\mathbf{b})(\mathbf{Z} - \mathbf{Xb})$$

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

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

$$W(\mathbf{b}) = \begin{pmatrix} \delta_1 & w_{12}(\mathbf{b}) & w_{13}(\mathbf{b}) & \dots & w_{1n}(\mathbf{b}) \\ 0 & \delta_2 & w_{23}(\mathbf{b}) & \dots & w_{2n}(\mathbf{b}) \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ & & & \ddots & \\ 0 & 0 & 0 & & w_{(n-1)n}(\mathbf{b}) \\ 0 & 0 & 0 & \dots & \delta_n \end{pmatrix}$$

- and

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

- In multivariate censored regression, the iteration concept is still applied to develop the Buckley-James estimators:

$$b_{n+1} = (X^T X)^{-1} X^T (Xb_n + W(b_n)(Z - Xb_n))$$

- Nevertheless if the iteration fails to converge, one can solve this problem by taking the average of all possible solutions of β

Previous Diagnostics

Renovated Diagnostics	Main purpose
Scatterplot (Smith & Zhang, 1995)	Describe relationship
Hat Matrix (Smith & Zhang, 1995)	Identify outliers
Added Variable Plot (Smith & Peiris, 1999)	Checking linearity
	Indicates new variable effects
Partial Residual Plot (Wang, Smith & Aziz, 2009)	Independent variable transformations
Cook's Distance (Aziz & Wang, 2009)	Discover influential observations

Local Influence Diagnostics

- Local Influence?

  Local influence was proposed by Cook in 1986 and it can be used to discover influential observations in a data set.
- How does it work?
  - The general influence function of $T \in R^{p+1}$, can be displayed as

$$GIF(T,h) = \lim_{\varepsilon \to 0} \frac{T(w_o + \epsilon h) - T(w_o)}{\epsilon}$$

    where $w = w_o + \epsilon h \in R^n$ describes a perturbation with the null perturbation, w_o fulfils $T(w_o) = T$ and $h \in R^n$ refers to a unit-length vector.
  - Next, one can specify generalised Cook statistics to measure the influence of the perturbations on T as

$$GC(T,h) = \frac{\{GIF(T,h)\}^T M \{GIF(T,h)\}}{c},$$

    where M is a $p \times p$ positive-definite matrix and c is a scalar.

Outline

Continue...

- Continue...How does it work?
  - One may find a direction of $h_{max}(T)$ to perturb a datum and maximize local change in T.
  - The direction of $h_{max}(T)$ can be derived by maximizing the absolute value of $GC(T, h)$ with respect to h.
  - The serious local influence appears if maximum value $GC_{max}(T) = GC(T, h_{max}(T))$.

Outline

Perturbing the variance

- By using the Buckley-James estimators as follows

$$b = (X^T QX)^{-1} X^T QY^* \tag{2}$$

perturb the variance of the error in (2), by replacing ϵ as $\epsilon_w \sim N(0, \sigma^2 W^{-1})$.

- Let W be diagonal matrix

$$W = \begin{pmatrix} w_1 & 0 & \cdots & 0 \\ 0 & w_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & w_n \end{pmatrix}$$

and vector $w^T = (w_1, w_2, \ldots, w_n)$ and w is given by $w = w_\circ + \epsilon h$, where $w_\circ^T = (1, 1, \ldots, 1)$, the n-vector of ones and $h^T = (h_1, h_2, \ldots, h_n)$ refers to a unit-length vector.

- Hence, W can be written as

$$W = \mathbf{I}_n + \epsilon D(h), \tag{3}$$

where $I_n = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}$ and $D(h) = \begin{pmatrix} h_1 & 0 & \cdots & 0 \\ 0 & h_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & h_n \end{pmatrix}$.

Outline

Continue...Perturbing the variance

- Now (2) becomes
$$b(w) = (X^T WQX)^{-1} X^T WQY^*. \tag{4}$$
- By replacing $W = \text{diag}(w_1, w_2, \ldots, w_n)$ in (4), $b(w)$ can be rewritten as below
$b(w) = [(X^T QX)^{-1} - \epsilon \left\{ (X^T QX)^{-1} X^T QD(h) X (X^T QX)^{-1} \right\}] \times X^T WQY^*$, where
$X^T WQY^* = X^T \{\mathbf{I}_n + \epsilon D(h)\} QY^* = X^T QY^* + \epsilon X^T QD(h) Y^*$.
- Therefore, $b(w)$ is given by
$$b(w) = b + \epsilon \left\{ (X^T QX)^{-1} (X^T QD(h) e^*) \right\} + O(\epsilon^2). \tag{5}$$
where $e^* = Y^* - Xb$.From (5), the $GIF(b, h) = (X^T QX)^{-1} X^T QD(e^*)h$.
- Next, the generalised Cook statistic of b is developed. It is scaled by $M = X^T \triangle X$, following that $cov(b) = (X^T \triangle X)^{-1} \sigma^2_{BJ}$, where $\triangle = \text{diag}(\delta_1, \delta_2, \ldots, \delta_n)$. Therefore
$$GC_1(b, h) = \frac{h^T D(e^*)(H^*)^2 \triangle D(e^*)h}{ps^2}, \tag{6}$$
where $H^* = X(X^T QX)^{-1} X^T Q$ and s^2 is the estimate variance.
- By applying $M = X^T X$ to the scaled generalised Cook statistic, which is based on $cov(b) = (X^T X)^{-1} \sigma^2$,
$$GC_2(b, h) = \frac{h^T D(e^*)(H^*)^2 D(e^*)h}{ps^2}. \tag{7}$$

Outline

Perturbing the response variables

- The response variable can be perturbed as $Y_w^* = Y^* + \varepsilon h$, where $h \in R^n$ refers to a unit-length vector. Let equation $(X^T QX)^{-1} X^T QY^*$ become

$$(X^T QX)^{-1} X^T QY_w^* = b + \varepsilon (X^T QX)^{-1} X^T Qh. \tag{8}$$

- Therefore, the general influence function of b under the perturbation can be shown as

$$GIF(b, h) = (X^T QX)^{-1} X^T Qh. \tag{9}$$

- Now two generalised Cook statistics can be developed by using the scale $M = X^T \triangle X$ and $M = X^T X$, which are

$$cov(b) = \begin{cases} (X^T \triangle X)^{-1} \sigma_{BJ}^2 & \text{if (censored regression),} \\ (X^T X)^{-1} \sigma^2 & \text{if (LSR).} \end{cases} \tag{10}$$

- Hence, $GC_1(b, h) = \dfrac{h^T (H^*)^2 \triangle h}{ps^2}$ and $GC_2(b, h) = \dfrac{h^T (H^*)^2 h}{ps^2}$.

Outline

Perturbing independent variables

- If one perturbs the ith column of X as $X_w = X + \epsilon l_i h d_i^T$,
- $(X_w^T Q X_w)^{-1} = (X^T Q X)^{-1} - \epsilon l_i (X^T Q X)^{-1} \times (X^T Q h d_i^T + d_i h^T Q X + d_i h^T h d_i^T)(X^T Q X)^{-1} + O(\epsilon^2)$ and $X_w^T Q Y^* = (X + \epsilon l_i h d_i^T)^T Q Y^* = X^T Q Y^* + \epsilon l_i d_i h^T Q Y^*$.
- Later, $(X_w^T Q X_w)^{-1}(X_w^T Q Y^*) = b + \epsilon l_i (X^T Q X)^{-1} \left\{ d_i h^T Q(e^*) - X^T Q h d_i^T b \right\} + O(\epsilon^2)$.
- Thus the general influence function of b, $GIF(b,h) = l_i (X^T Q X)^{-1}[d_i h^T Q(e^*) - X^T Q h d_i^T b]$. Replace the *ith* element of b, therefore $d_i^T b = b_i$ and $GIF(b,h) = l_i (X^T Q X)^{-1}[d_i (e^*)^T - b_i X^T] Q h$.
- Then two generalised Cook statistics for b are constructed as

$$GC_1(b,h) = \frac{l_i^2 h^T H^* \triangle \left\{ e^* d_i^T - b_i X \right\} (X^T Q X)^{-1} \left\{ d_i (e^*)^T - b_i X^T \right\} Q h}{p s^2} \text{ and}$$

$$GC_2(b,h) = \frac{l_i^2 h^T H^* \left\{ e^* d_i^T - b_i X \right\} (X^T Q X)^{-1} \left\{ d_i (e^*)^T - b_i X^T \right\} Q h}{p s^2}.$$

- One can obtain the diagnostic direction h_{max} by computing the eigenvector corresponding to the largest eigenvalue of the following matrice $H^* \triangle \left\{ e^* d_i^T - b_i X \right\} (X^T Q X)^{-1} \left\{ d_i (e^*)^T - b_i X^T \right\} Q$, or $H^* \left\{ e^* d_i^T - b_i X \right\} (X^T Q X)^{-1} \left\{ d_i (e^*)^T - b_i X^T \right\} Q$

Outline

Illustration

- Data: Stanford heart transplant data
- n=152 patients, 55 deceased ($\delta_i = 1$) while 97 are still alive ($\delta_i = 0$)
- Explanatory variables = censored status, age at time of first transplant (in years) and T5 mismatch score.
- The response variable = survival time(days)
- The Buckley-James model for this data set was developed as $Y = \beta_0 + \beta_1 AGE + \beta_2 AGE^2 + \beta_3 T5$.
- First, consider the variance perturbation. The index plot of $|h_{max}|$ in the next slide shows patients aged below 20 years as the most influential cases.
- This finding agrees well with Reid and Crepeau (1985), and Pettitt and Daud (1989)

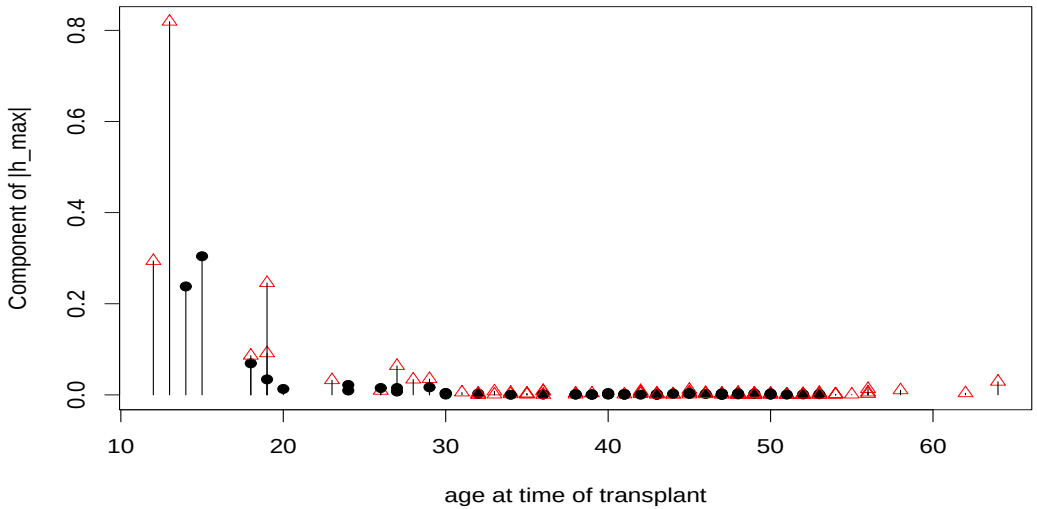

Figure: Index plots of $|h_{max}|$ for perturbing variance for Stanford heart transplant data ($n = 152$).

Outline

Continue...Illustration

- Next, consider the perturbation of response variable and individual independent variables.
- It is obvious that the most influential patients are aged below 20 years and two patients aged above 60 years.
- Removal of the patients aged 12 and 13 decreases $\hat{\beta}_1$ by 0.010 and 0.030 respectively, while removal of the patient aged 15 increases $\hat{\beta}_1$ by 0.015
- There is no impact on the estimator values in the Buckley-James model when deleting those observations (one at a time) since the maximum eigenvalues for the perturbation of the variance, response variable, x_1 and x_2 are small at 0.142, 0.021, -0.002 and 1.000 respectively.
- However, when the p-value is scrutinized, one can find the p-value for x_1 is roughly five times larger when deleting case 1, and triple when deleting case 4, whereas deleting case 2 has a large effect on the p-value of x_2 where the value becomes fourteen times larger.
- No attention is given to x_3 since this variable is not strongly associated with survival time.

Conclusion

- The proposed local influence diagnostics for the Buckley-James model performs very well for identifying influential cases and for assessing the effects that perturbations to the assumed data would have on inferences.
- It should also be noted that the proposed diagnostics is able to easily detects influential observations from both groups i.e. censored and uncensored observations in the data set as opposed to the previous diagnostics for Buckley-James model.

Thank you