Statistika, Vol. 8 No. 2, 103 – 109
Nopember 2008

A Detection Measure of Outliers Based on Forward Search Approach for Cox-Regression Model

NOR AKMAL MD NOH[1)], IBRAHIM MOHAMED[1)], AND NUR AISHAH MOHD TAIB[2)]

1) Institute of Mathematical Sciences, University of Malaya 50603 Kuala Lumpur, Malaysia
2) Department of Surgery, University of Malaya Medical Centre, 50603 Kuala Lumpur, Malaysia
Email: akmal82@perdana.um.edu.my, imohamed@um.edu.my, naisha@um.edu.my

ABSTRACT

This paper focuses on identifying possible outliers based on Cox regression model. Forward search method has been applied in several studies involving regression-based models such as linear regression and generalized linear model. The method starts with a pre-selected subset of a data set. The method moves forward through the data by adding observations one by one and progressive changes in values of statistics are noted. In this paper, we extend the application of forward search in survival data analysis. Currently, graphical methods are used to detect any significant changes in values of the statistics. We propose a measure which may aid us in determining observations that are outlier.

Keywords: *Outliers, Forward Search, Cox Regression Model.*

1. Introduction

It is well known that the existence of the single outlier may alter the parameter estimation of any fitted models in all areas of research. Belsley et al. (1980), Barnett and Lewis (1984) and Montgomery and Peck (1992) had discussed the topics for linear regression problems. However, the research on detection on influential observations or outliers in survival data analysis has not received enough attention. Several survival models including Cox's regression model and parametric models based on common distributions such as exponential and Weibull distributions are widely used in survival study (see Cox, 1972 and Grambsch & Therneau, 1994). It is known that outliers in survival data set will affect the parameter estimation of the models and consequently alter the hazard ratios. Thus, it is important to determine any observation that affect the inference based on the fitted model to an observed set of survival data.

Outlier in survival study is defined slightly differently from outlier in linear regression problem. This is because the dependent variables in survival models contains the information on the survival times and status of patients in the study. Let $\hat{S}_i(t_i)$ denotes the estimated value of the survival function for i^{th} individual at his observed death time t_i. Nardi and Schemper (1999) stated that the prediction of survival probability by Cox's model is considered good for i^{th} individual if $\hat{S}_i(t_i)=0.5$. Alternatively, if the observed survival time and estimated median survival time based on the fitted Cox-regression model are close, then the data is fitted well by the model. Otherwise, we should perform further diagnosis checking including identification of outliers.

The common approach employed in detecting outlier is through residual analysis. In linear regression problem, various methods based on residual analysis including the leave-one-method (see Hadi (1992) and Imon (2005)) had successfully identified outliers in the problems considered. In survival analysis, the deviance residuals have become a standard type of residuals used for detection of the influential observations or outliers in survival model. Deviance residuals are symmetrically distributed around zero and have no reference sampling distribution with or without censoring. Nardi and Schemper (1999) had introduced two new types of residuals, the log-odds and normal deviate residuals, for similar purposes. On the other

hand, Atkinson and Riani (2000) had proposed an alternative method called forward search method for linear regression problems. Based on the residuals obtained from the fitted model, an initial subset of size *N* (*N* is smaller than the size of data set) is taken from the data. The effect of adding one observation into the initial subset at one time on the statistics of interests will be continuously monitored until all observations are in the subset. Including influential observation is expected to cause some significance changes in the estimates of the statistics. The approach has also been successfully applied to several models including regression and generalized linear models.

In this paper, we extend the use of forward search method to identify outlier in survival data. The pre-selected subset and the order of observation entering the subset are determined by the order of squared deviance residuals for every observation. Usually, the effects of observations entering the subset on the statistics of interest are observed through the progression plots. We suggest a simple statistics to identify observation which has large effect of the statistics of interest.

This paper is organized as follows. The theory of Cox's regression model and deviance residuals are reviewed in Section 2 and Section 3 respectively. Forward search method is described at length in Section 4. In Section 5, we define a new measure to identify influential observations. Lastly, in final section, we illustrate the application of the residuals on a prostate cancer data set used by Nardi and Schemper (1999).

2. Cox PH Model

In this paper, we consider the Cox proportional hazard model with covariates. The hazard of death at a particular time depends on the values $x_1, x_2 \cdots x_p$ of p explanatory variables $X_1, X_2 \cdots X_p$. Let $\boldsymbol{x} = (x_1, x_2 \cdots x_p)'$ and $h_0(t)$ be the hazard function for an individual with $\boldsymbol{x}_j = 0$, for $j = 1,2,..., p$ called the baseline hazard function. Then, the hazard function for the i^{th} individual is given by

$$h_i(t) = \psi(\boldsymbol{x}_j) h_0(t) \tag{2.1}$$

where, $\psi(\boldsymbol{x}_i)$ is a function of the values of the vector of explanatory variables of the i^{th} individual and ψ is interpreted as the hazard at time t for individual whose vector of explanatory variables is $\boldsymbol{x}_j$, relative to the hazard for an individual for whom $x_j = 0$. Since $\psi(\boldsymbol{x}_j)$ >0, then hazard function for the i^{th} individual is given by

$$h_i(t) = \exp(\eta_j) h_0(t) \tag{2.2}$$

where

$$\eta_j = \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi} \tag{2.3}$$

is called the linear component of the model, or the risk score, or the prognostic index for the i^{th} individual. Model (2.2) can be written as

$$\log\left\{\frac{h_i(t)}{h_0(t)}\right\} = \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi} \tag{2.4}$$

Now, the proportional hazard model is regarded as a linear model for the logarithm of the hazard ratio.

3. Deviance Residuals

Deviance residuals were introduced by Therneau *et al.* (1990). The residuals are closely symmetrically distributed about zero but have no reference on sampling distributions. It is usually used in medical research on detecting abnormality observations in the survival model. They are defined by

$$r_{D_i} = \operatorname{sgn}\left(r_{M_i}\right)\left[-2\left\{r_{M_i} + \delta_i \log\left(\delta_i - r_{M_i}\right)\right\}\right]^{\frac{1}{2}} \tag{3.1}$$

where $r_{M_i} = \delta_i - \exp(\hat{\beta}' x_i)\hat{H}_o(t_i)$, the martingale residuals

$\delta_i = 0$ for the censored observations and $\delta_i = 1$ for the uncensored observations

$i = 1,2...,n$ and n are number of individuals in sample.

The second term in r_{M_i} is an estimate of $\hat{H}_i(t_i)$, the cumulative hazard or cumulative probability of death of the i^{th} individuals over the interval $(0,t_i)$. On the other hand, $\hat{H}_o(t_i)$ is known as the cumulative baseline hazard function given by $\hat{H}_0(t) = -\sum_{j=1}^{k} \log \xi_j$, and $\tilde{\xi}_j = 1 - \dfrac{d_j}{\sum l \in R(t_{(j)}) \exp(\hat{\beta} \underline{x} l)}$.

4. Forward Search Method

The forward search method is developed based on the stepwise method found in regression theory. It has been applied in several research areas including generalized linear model (see Atkinson and Riani, 2000). The method monitors the effects of including new observation into a data set one at a time on the parameter estimates and other statistics of interests of the models. For survival analysis, the method comprises three main steps described below:

First step: Choosing the Initial Subset or 'clean' data set.

The full data set is fitted using the Cox-regression model (2.4) and the deviance residuals, r_{D_i}, are obtained. We rank the observations based on the squared deviance residuals $r_{D_i}^2$. The initial subset of data is formed by $\text{median}\left(\beta^{(-j)}\right) \pm 3\text{MAD}\left(\beta^{(-j)}\right)$ of the observations, say of size k. Denote the initial subset as $S_*^{(m)}$.

Second step: Adding observation

The next step is to choose an observation to be included into $S_*^{(m)}$. We fit model (2.4) on the $S_*^{(m)}$ giving the parameter estimates, $\hat{\beta}_m^{(-j)}$, $j = 1,2,...,p$, and residual variance, σ_m^2. We then use $\hat{\beta}_m^{(-j)}$ on the full data set to obtain the squared deviance residuals $r_{D_i}^2$. After ranking the $r_{D_i}^2$, the new subset $S_*^{(m+1)}$ of size $m+1$ is formed by $m+1$ observations with the smallest squared deviance residuals.

Third step: Repeat the process

We repeat the process in second step until we have $S_*^{(n)}$ that is all observations are in the subset.

Then, any changes on $\hat{\beta}_m^{(-j)}$ and σ_m^2 for $m = k,k+1,....,n$ can be observed visually using the appropriate plot such as line plot.

4.1. A New Measure of outlier Effect

We may monitor the values of parameter estimates at every steps of forward search method through graphical methods. However, it is quite difficult to identify the significance changes visually from the plots. It can be done easier if we have a statistics to measure the changes caused by the effect of including a new observation on the parameter estimates and variance directly. Let $\hat{\beta}_m^{(-j)}$ be the statistics of interest at step m of forward search method, $m = k,k+1,....,n$ and $j = 1,2,..., p$. Define effect of including an observation at m-step as

$$\mathrm{EM}_{\mathrm{m}} = \left|\beta^{(m)} - \beta^{(m-1)}\right| \tag{4.1}$$

We introduce the cut point of the statistics based on

$$\mathrm{EM}_{\mathrm{cp}} = \mathrm{mean}\left(\mathrm{EM}_{\mathrm{mj}}\right) + 2\mathrm{s.d}\left(\mathrm{EM}_{\mathrm{m}}\right) \tag{4.2}$$

Any EM_{m} that lies above EM_{cp} suggests that the observation that enters the subset at the j^{th} order time is a candidate to be an outlier. Similar approach is used for monitoring changes in σ_j^2 for $m = k,k+1,....,n$.

5. Data analysis

We consider the prostate cancer data from Andrews and Herzberg (1985) to illustrate the application of the proposed statistics. Only 310 observations are considered consisting patients of age below 75 years old. The covariates for the study are summarized as follows: treatment (treat; <0.2 mg diethylstilbestrol versus >0.2 mg), weight index (wt;<100, >100), performance rating (pf; normal, limitation of activity), serum hemoglobin (HG; <12 g/ml, >12 g/ml), size of primary lesion (sz; <30 cm2, >30 cm2), Gleason stage/grade category (SG; <10, >10), history of cardiovascular disease (HX; no, yes). Summary of statistics for the data is given in Table 1.

Table 1. Summary statistics for prostate cancer data

Statistics	Overall	Uncensored group	Censored group
Minimum	0	0	51
Mean	38.88	26.83	62.75
Median	39.5	24	64
Maximum	76	74	76
1st quartile	18	11	57
3rd quartile	59.75	39	68
Standard deviation	23.20	18.65	7.51

We perform model selection based on Cox-regression model (2.4) to the data set. We find out that 5 factors are significant; sg, HX, sz, rx, and wt. Based on the *cox.zph* procedure

available in *SPlus* the overall proportional hazard assumption for the model is satisfied (p-value = 0.870).

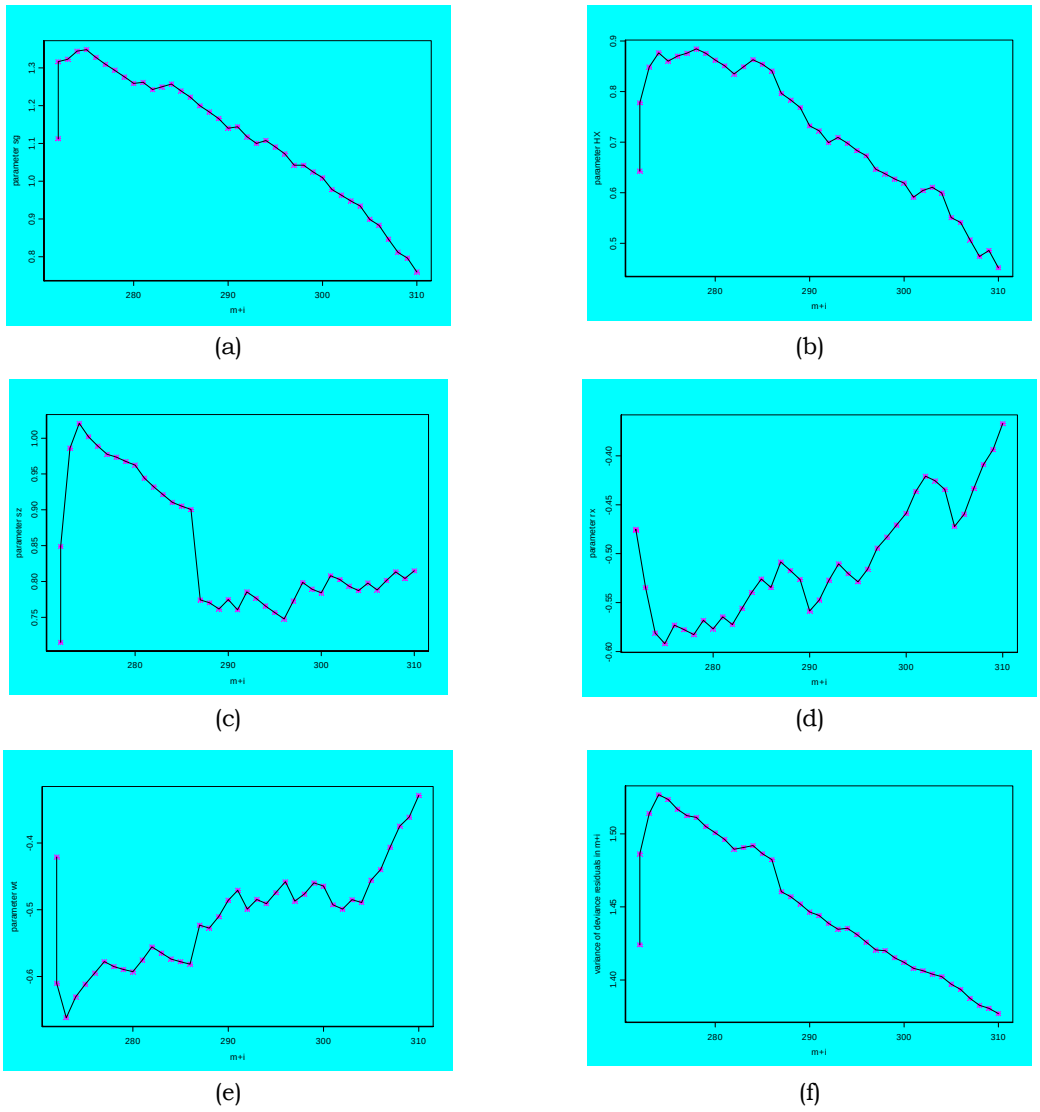

Figure 1. The Progression Plots Based on Forward Search Method

Table 2. List of Possible Outliers

Variable	Possible Outlier	Variable	Possible Outlier	Variable	Possible Outlier
Treatment	155,451	Size	135	history	135,451
Weight	135	Gleason	No outliers	Variance	135

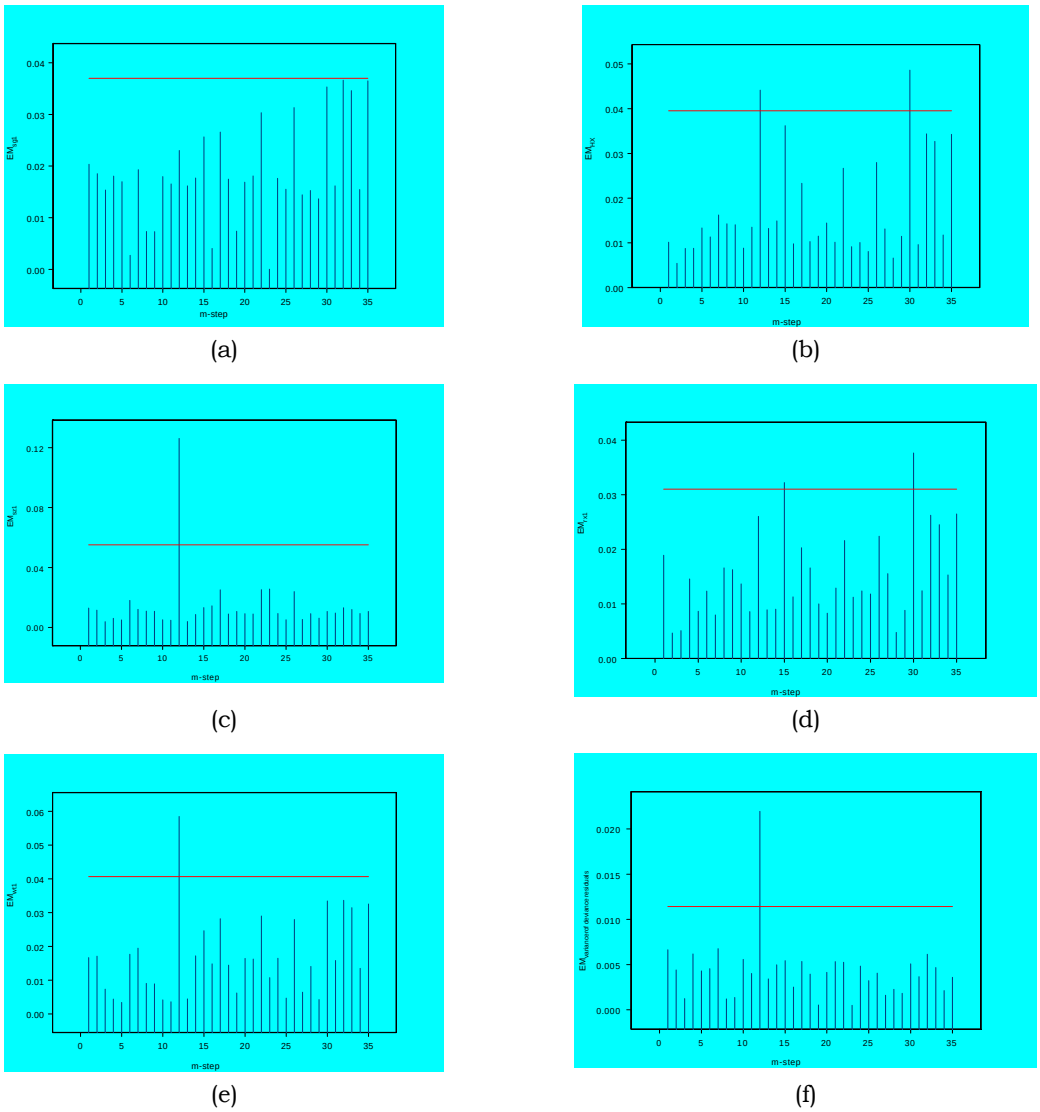

Figure 2. The Index Plots Based on Statistics EM

We apply the forward search method on the data with the initial subset was formed by $\text{median}\left(\beta^{(-j)}\right) \pm 3\text{MAD}\left(\beta^{(-j)}\right)$ of the original data size. The changes on parameter estimates are monitored throughout the process and plotted in Figure 1(a)-(e) while the changes on the variance are given in Figure 1(f). The line plots of measures EM are given in Figure 2. It can be seen that the identification of observation which changes the estimates most can be done easily from the plots of EM statistics. Possible list of outliers a given in Table 2.

Table 3. Background of patients identified as possible influential observations

Patient number	time	Status	rx1	wt1	sz1	sg1	HX
135	6	1	0	0	0	0	0
155	5	1	1	1	0	1	0
451	4	1	1	1	0	0	0

Table 3 gives the observations which show large changes and consequently are candidates of being outliers. Observation 135 gives large EM for almost every parameters and variance of residuals. The patient belong to "*no risk*" group as the value of variable are 2nd level, where he dies for early at time t = 6. Other observations that are worth to look at are observation 155 and 451. Generally, they are associated to patients who "die too early" case. Note that the survival times for all patients 135, 155 and 451 are well below the mean survival times for overall patients or uncensored groupings.

6. Conclusion

In this paper, we have considered the application of forward search method in identifying outliers in survival data set. We propose a simple measure to facilitate the identification process and apply on a prostate cancer data. The whole procedure may be improved further by selecting a more appropriate initial data set and measure of influential observations.

References

[1]. D.F., Andrews, and A.M., Herzberg. (1985). *Data*. Springer, New York.
[2]. A.C., Atkinson, and M., Riani. (2000). *Robust Diagnostic Regression Analysis*, Springer, New York.
[3]. V., Barnett, and T., Lewis. (1984). *Outliers in Statistical Data*. Second edition. Wiley, London.
[4]. D.A., Belsley, E., Kuh, and R.E., Welsch. (1980). *Regression diagnostics: Identifying influential data and sources of collinearity*. Wiley, New York.
[5]. D.R., Cox. (1972). Regression model and life table (with Discussion). *Journal of the Royal Statistical Society*, Series B, 34, 187-220.
[6]. P.M., Grambsch, and T.M., Therneau. (1994). Proportional hazards test and diagnostics based on weighted residuals. *Biometrics*, 81, 515-526.
[7]. A.S., Hadi, A.S. (1992). A new measure of overall potential influence in linear regression. *Computational Statistics & Data Analysis*, 14, 1-27.
[8]. A.H.M., Imon. (2005). A stepwise procedure for the identification of multiple outliers and high leverage points in linear regression. *Pak. J. Statist.*, 21(1), 71-86.
[9]. D. C. Montgomery, and E. A. Peck, (1992). *Introduction to linear regression analysis*, second edition, Wiley, New York.
[10]. A., Nardi and M., Schemper. (1999). New Residuals for cox regression and their application to outliers screening, *Biometrics*, 55, 523-529.
[11]. T.M., Therneau, P.M., Grambsch, and T., Flemming. (1990). Martingale based residuals for survival models. *Biometrika*, 77, 147-160.