Austrian Journal of Statistics
January 2023, Volume 52, 66–85.
http://www.ajs.or.at/
doi:10.17713/ajs.v52i2.1393

Bootstrap Based Diagnostics for Survival Regression Model with Interval and Right-Censored Data

Jayanthi Arasan
Universiti Putra Malaysia

Habshah Midi
Universiti Putra Malaysia

Abstract

This research proposes a new approach based on the bias-corrected bootstrap harmonic mean and random imputation technique to obtain the adjusted residuals (Hboot) when a survival model is fit to right- and interval-censored data with covariates. Following that, the model adequacy and influence diagnostics based on these adjusted residuals, case deletion diagnostics, and the normal curvature are discussed. Simulation studies were conducted to assess the performance of the parameter estimate and compare the performances of the traditional Cox-Snell (CS), modified Cox-Snell (MCS) and Hboot at various censoring proportions (cp) and samples sizes (n) using the log-logistic and extreme minimum value regression models with right- and interval-censored data. The results clearly indicated that Hboot outperformed other residuals at all levels of cp and n, for both models. The proposed methods are then illustrated using real data set from the COM breast cancer data. The results indicate that the proposed methods work well to address model adequacy and identify potentially influential observations in the data set.

Keywords: bootstrap, residual, influence, interval-censored, harmonic.

1. Introduction

The problem of mixed case interval-censored data is very common in medical research where inspections on patients are conducted at different time intervals. So, lifetime of the i^{th} subject is only known to fall within an interval, $L_i < t_i < R_i$, where L_i and R_i are known as left and right endpoints. The subject is right-censored when $L_i < t_i < \infty$ or, the subject has been event-free at the last known time L_i. So right-censored data is a special case of interval-censored data, see Sun (2006) and Kalbfleisch and Prentice (2011).

One of the most widely used residual in survival analysis is known as the Cox-Snell residual which is given by $r_{C_i} = \hat{H}(t_i) = -\log(\hat{S}(t_i))$ for the i^{th} subject, $i = 1, 2, \cdots, n$, where $\hat{H}(t_i)$ and $\hat{S}(t_i)$ are the estimated cumulative hazard and survivor functions respectively. The usefulness of this residual arises from the fact that the distribution of $-\log(S(t))$ follows the exponential distribution with unit mean, exp(1), see Collett (2015). Thus, whenever a model is fit to any survival data, if $-\log(\hat{S}(t_i))$ follows the unit exponential distribution closely, we can be assured that an apt model has been fit to the data.

Researchers have suggested adjustments to the Cox-Snell residuals to make up for the excess,

ζ, in the residuals when data is right-censored. Two traditional modifications of r_{C_i} take ζ as the mean and median of the unit exponential distribution, which is 1 and 0.693, respectively, see Crowley and Hu (1977). Recent works by Naslina, Jayanthi, Syahida, and Bakri (2020) and Lai and Arasan (2020) have shown that the modified residuals that are based on the empirical geometric and harmonic mean outperform traditionally used residuals. When data is interval-censored, the Cox-Snell residuals are themselves intervals censored and Farrington (2000) suggests replacing the interval residuals with expected values under exp(1).

The bootstrap method as discussed by Efron and Tibshirani (1994) is a direct application of the plug-in principle which is a way of understanding the population, based on estimates from random samples drawn from the population. Alternative bootstrap-based estimates and techniques have been showing promising results when dealing with moderately censored data as discussed by Arasan and Lunn (2009, 2008), Manoharan, Arasan, Midi, and Adam (2017), and Meeker, Escobar, and Pascual (2014). The theoretical aspect of the single bootstrap was studied by Singh (1981) and, Abramovitch and Singh (1985). More recent works are by Zhuang, Xu, and Pang (2021) who proposed an improved two-stage method combined with fractional-random-weight bootstrap to analyze interval failure data, and Grzegorzewski, Hryniewicz, and Romaniuk (2020) who proposed a new methodology for simulating bootstrap samples of fuzzy numbers.

Some earlier works on influence diagnostics for survival models are by authors such as Pettitt and Daud (1989), Escobar and Meeker Jr (1992), Weissfeld and Schneider (1990), Lesaffre and Verbeke (1998) and Ortega, Cancho, and Bolfarine (2006). There are not many research works that discuss influence diagnostics for survival models with interval-censored data. Recently, Manoharan, Arasan, Midi, and Adam (2020) assessed the performance of the local influential diagnostics of the extended log-normal distribution with left-truncation and interval-censored data.

This research proposes two new adjustments to the traditionally used residuals in survival analysis when the lifetimes are right and interval censored. The first adjustment estimates the lifetime for each interval censored observation using the mean of $R = 1000$ randomly imputed lifetimes based on the estimated left and right survivor functions. The second adjustment uses the bias-corrected bootstrap harmonic mean to address the excess in the Cox-Snell residual when data is right censored because bootstrap estimates are obtained directly from the data and may perform better than estimates that are based on theoretical assumptions. The harmonic mean is employed instead to the arithmetic mean because it is known to better handle extreme values and outliers.

2. Model adequacy and influence diagnostics

2.1. Model with right and interval-censored data

Let $S(t, \boldsymbol{\beta}|\boldsymbol{x})$ denote the survival distribution of a non-negative continuous random variable, T with parameters $\boldsymbol{\beta}$ given the covariates $\boldsymbol{x}$. Suppose we have both interval-censored and right-censored lifetimes for $i = 1, 2, \ldots, n$ observations. Let L_i and R_i be the left and right endpoints for the i^{th} subject. The following indicator variable is used to identify whether the i^{th} observation is interval or right-censored.

$$\delta_i = \begin{cases} 1 & \text{for } t_i \text{ interval-censored,} \\ 0 & \text{for } t_i \text{ right-censored.} \end{cases}$$

The log-likelihood function of the full sample with interval and right-censored data is,

$$L(\beta, \sigma) = \prod_{i=1}^{n} [S(L_i) - S(R_i)]^{\delta_i} [S(L_i)]^{(1-\delta_i)} . \tag{1}$$

2.2. Adjusted Cox-Snell residual for interval and right-censored data

In this section, we discuss methods that can be used to check the model adequacy and detect influential observation and outliers when a survival model is fitted to right and interval-censored data. We propose methods that incorporate the use of bias-corrected bootstrap estimates and a random imputation technique to assess the model adequacy and carry out influence diagnostics. The residuals proposed by Farrington (2000) for interval-censored data are themselves interval-censored and he tried to overcome this problem by replacing the interval residuals with expected values under exp(1). Thus, it may not be that feasible when we are dealing with more complex survival models. Many researchers employ the midpoint imputation technique to deal with the problem of interval-censored data, see Lui, Darrow, and Rutherford III (1988); Mariotto, Mariotti, Pezzotti, Rezza, and Verdecchia (1992). Midpoint imputation is not that efficient because it only uses two endpoints to approximate the actual failure time, which is highly unreliable when the interval length is wide, see Law and Brookmeyer (1992).

To better deal with the problem of interval-censored data, we propose generating $R = 1000$ random variates from $U(\hat{S}(R_i), \hat{S}(L_i))$ for the i^{th} observation where $\hat{S}(.)$ represents the estimated survivor function of the proposed model for the data. Following that we transform each of these survival probabilities to obtain the 1000 replication of each estimated lifetimes, ${t_i}^r$, where $r = 1, 2, \cdots, R$. Then, the estimated lifetime for the i^{th} observation is $t_i^* = \left(\sum_{r=1}^{R} \frac{{t_i}^r}{R}\right)$, and the adjusted Cox-Snell residual when data is interval-censored is,

$$r_{C_i}^{\mathrm{I}} = \hat{H}(t_i^*) = -\log(\hat{S}(t_i^*)).$$

Existing adjustments to the Cox-Snell residuals for right-censored data may not work well when mixed censoring schemes are present in the data, leading to positively skewed data due to extreme values or outliers. In this research, we propose adjusting the Cox-Snell residuals using the bias-corrected bootstrap harmonic mean. Residuals based on these bootstrap estimates relieve us from making assumptions and having to depend solely on the traditional methods derived from statistical theory. They are obtained directly from the data and may perform better than estimates that are based on theoretical assumptions.

The harmonic mean, described by the reciprocals of the data is known to better handle extreme values and outliers because it is much lesser affected by them. Thus, when data is moderately or heavily censored, we do expect the harmonic mean to better represent the average of the excess residuals compared to other traditional estimates as discussed by Naslina *et al.* (2020) and Lai and Arasan (2020). The bootstrap estimate of bias requires the sampling of a large number, B, of bootstrap samples with replacement from the original Cox-Snell residuals, with each observation having an equal probability of being chosen. Then, $h^*_{(b)}$, $b = 1, 2, \cdots, B$ are the sample harmonic means calculated from each of these bootstrap samples of size n. The estimated bootstrap bias for the harmonic mean is given by, $h^*_{(.)} - h$, where h is the harmonic mean calculated from the original Cox-Snell residuals. Following that, the bias-corrected bootstrap harmonic mean can be obtained as follows:

$$h_{bc} = 2h - h^*_{(.)}, \tag{2}$$

where,

$$h^*_{(.)} = \frac{\sum\limits_{b=1}^{b} h^*_{(b)}}{B}.$$

The adjusted Cox-Snell residuals when data is right-censored is,

$$r^{\mathrm{R}}_{C_i} = r_{C_i} + h_{bc}, \quad \text{for the } i^{\text{th}} \text{ subject},$$

where $r_{C_i} = -\log(\hat{S}(t_i))$ are the Cox-Snell residuals obtained using the imputed lifetimes. Following that, the adjusted Cox-Snell residuals when data is right and interval-censored is,

$$r^*_{c_i} = \begin{cases} r^{\mathrm{R}}_{C_i} & \text{when data is right-censored and,} \\ r^{\mathrm{I}}_{C_i} & \text{when data is interval-censored.} \end{cases}$$

All discussions regarding the residuals should also hold for the adjusted residuals. Thus, the plot of $\log[-\log(\hat{S}(r^*_{C_i}))]$ against $\log(r^*_{C_i})$ can be constructed to check if an adequate model has been fit to the data, where $\hat{S}(r^*_{C_i})$ is the estimated Kaplan-Meier (KM) survivor function based on the values of $r^*_{C_i}$. In this case, the plot should reveal a straight line with a unit slope and 0 intercept. Outliers can also be revealed by this plot as it would stand out from the rest of the observations due to a very high residual value.

2.3. Adjusted martingale residuals

The well-known martingale residual was another adjustment to the Cox-Snell residual intended to relocate the mean of the Cox-Snell residual from unity to 0 when the observations are uncensored. It was first discussed by Lagakos (1981) and then by Barlow and Prentice (1988) and Therneau, Grambsch, and Fleming (1990). The adjusted Cox-Snell residuals can be used to obtain the adjusted martingale residuals as given by the following,

$$r^*_{M_i} = \delta_i - r^*_{C_i}.$$

The plot of $r^*_{M_i}$ versus observation number can show observations that are not well fitted by the model, also known as outliers. Plots versus the covariates can also be very useful in checking poorly fit subjects. However, because the martingale residuals tend to be asymmetric the deviance residuals will in general be more powerful in detecting outlying subjects.

2.4. Adjusted deviance residuals

The deviance residual was introduced by Therneau *et al.* (1990) to adjust the martingale residuals so that they are more symmetrically distributed around 0, and thus easier to interpret. The modified deviance residuals can thus be obtained using the adjusted martingale residuals as given by the following,

$$r^*_{D_i} = \mathrm{Sgn}(r^*_{M_i})[-2(r^*_{M_i} + \delta_i \ln(\delta_i - r^*_{M_i}))]^{1/2},$$

where Sgn() is the sign function that takes the value of +1 and −1 for positive and negative arguments respectively. The plot of r^*_{Di} versus observation number is very effective at indicating outliers. Plot versus the covariates can also be constructed to check for any obvious pattern, and possible extreme observations and outliers.

2.5. Adjusted score residuals

The score or Schoenfeld residual was proposed by Schoenfeld (1982), and are derived from are the components of the first derivatives of the log-likelihood function with respect to its parameters. Thus, they take different sets of values for each parameter in the model. The adjusted score residuals (r^*_{Si}) when data is interval censored can be computed using the imputed lifetimes discussed in Section 2.2. The plot of r^*_{Si} versus observation number should be randomly distributed around zero for a good fit. Index plots of the score residuals for each covariate in the fitted model would be useful at indicating extreme observations and outliers.

2.6. Influential diagnostic

Influential observations are observations that have a significant amount of influence on the estimated parameters, model fit, and covariates as discussed by Hosmer, Lemeshow, and May (2008). The local influential diagnostics introduced by Cook (1986), identifies influential observations by introducing small perturbations to the data using weights without the need to completely remove potentially influential observations from a data set. In this section, we propose conducting the influence diagnostics using the imputed lifetime discussed in Section 2.2 when data is interval-censored.

Overall influence

The easiest method to assess the influence of observations on the estimated parameters is by using case-deletion diagnostics. This is carried out by computing the difference in the MLEs, $\Delta_i\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_{(i)}$, where $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}}_{(i)}$ are the MLEs calculated using full sample and sample with the i^{th} term deleted, respectively. The $\Delta_i\hat{\boldsymbol{\theta}}$ values are usually referred to as delta-beta or DFBETA.

Standardized delta-beta values are referred to as DFBETAS and the index plot of either DFBETA or DFBETAS for each parameter in the model can then be used to detect influential observations. Pettitt and Daud (1989) cautioned that using single-case deletion statistics may cause a masking effect, where the removed observation may mask the effect of a real outlier. They suggested studying the change caused by the perturbation to the likelihood displacement statistics.

Local influence statistics

The measures of curvature as discussed by Cook (1986) and Pettitt and Daud (1989) can be used to assess influence by examining the effect of perturbations to the data or covariates, using appropriate weights, $\boldsymbol{\omega} = (\omega_1, \omega_2, \cdots, \omega_n)$ which is the $n \times 1$ vector of perturbations restricted to some open subset $\boldsymbol{\Omega} \subset \mathbb{R}^n$. Here, $0 < \omega_i < 1$ and $\boldsymbol{\omega}_0 = (1, 1, ..., 1)^T$ is the vector of no perturbation and $L(\boldsymbol{\theta}|\boldsymbol{\omega}_0) = L(\boldsymbol{\theta})$.

The normal curvature at direction d is defined by,

$$c_d = 2\left|d^T \Delta^T (I)^{-1} \Delta d\right|,$$

where $\|d\| = 1$, I is observed information matrix evaluated at $\hat{\boldsymbol{\theta}}$, Δ is the $(p{+}1){\times}n$ matrix with elements $\Delta_{ki} = \frac{\partial^2 l(\boldsymbol{\theta}|\boldsymbol{\omega})}{\partial\theta_i \partial\omega_k}$ evaluated at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$, $\boldsymbol{\omega} = \boldsymbol{\omega}_0$, $k = 1, 2, \cdots, p+1$ and $i = 1, 2, \cdots, n$. $l(\boldsymbol{\theta}|\boldsymbol{\omega})$ is the perturbed likelihood and $p+1$ denotes the number of parameters in the model. Following that large elements of the eigenvector $d_{\max}$ associated with the maximum curvature $C_{\max}$ would indicate potentially influential observations. Following if $\hat{\boldsymbol{\theta}}$ and $\hat{\theta}_{(\omega)}$ are the MLEs under $L(\boldsymbol{\theta})$ and $L(\boldsymbol{\theta}|\boldsymbol{\omega})$, the likelihood displacement statistic to assess the influence of $\boldsymbol{\omega}$, is given by the following,

$$LD(\boldsymbol{\omega}) = 2[\log L(\hat{\boldsymbol{\theta}}) - \log L(\hat{\boldsymbol{\theta}}_{(\omega)})].$$

The local influence tries characterizing the behaviour of $LD(\omega)$ around ω_0. So the index plot of $d_{\max}$ would help us identify observations that have significant amount of influence on $LD(\omega)$. The likelihood displacement statistics for a perturbation on the i^{th} case, where $\boldsymbol{\omega} = (0, 0, \cdots, 1, \cdots, 0)^T$ and 1 is in the i^{th} position can be approximated by $\frac{1}{2}\ddot{F_{ii}}$ where $\ddot{F_{ii}}$ is the diagonal element of $\Delta^T(I)^{-1}\Delta$, see Escobar and Meeker Jr (1992). The index plot of $\frac{1}{2}\ddot{F_{ii}}$ can be used as a warning of a significant overall influence on $\boldsymbol{\theta}$ if the values exceed $\frac{1}{2}\chi^2_{(0.5;p+1)}$.

3. Simulation study

Two simulation studies were conducted via the R programming language using 1000 replications, at sample sizes (n) of 50 and 120, and approximate right censoring proportions (cp) of $0.30, 0.40, 0.50, 0.55$ and 0.65. Two survival models were assessed, the log-logistic and extreme minimum value regression models with a covariate. The first simulation study was conducted to assess the performance of the estimates based on the values of bias and mean square error (MSE).

If $z = \frac{y-\boldsymbol{\beta'x}}{\sigma}$, where $y = \log\, t$, the survivor functions of the log-logistic regression (LLR) and extreme minimum value regression (EMVR) distributions are,

$$S(y, \boldsymbol{\beta}, \boldsymbol{x}) = \frac{1}{1+e^z}, \quad -\infty < y < \infty,$$

$$S(y, \boldsymbol{\beta}, \boldsymbol{x}) = e^{-\exp(z)}, \quad -\infty < y < \infty,$$

respectively, $\boldsymbol{x'} = (x_0, x_1, \cdots, x_{p-1})$ is the vector of covariate values, where $x_0 = 1$ and $\boldsymbol{\beta'} = (\beta_0, \beta_1, \cdots, \beta_{p-1})$ are unknown parameters, $-\infty < \boldsymbol{\beta} < \infty$ and $\sigma > 0$.

The values for parameters β_0, β_1, and σ were chosen as 2, -0.8, and 0.5 for the LLR model and 2.58, -0.0136, and 1 for the EMVR model. These parameter values were chosen specifically to mimic survival data that are measured in months. The covariate values were simulated from the standard normal distribution. The survival times were obtained using the inverse transformation method. The censoring time for the i^{th} observation, $c_i \sim \exp(\mu)$, where the value of μ would be adjusted to obtain the desired approximate right censoring proportion in our data. In order to generate interval censored data, we use a sequence $\kappa = 24$ check-up times, $\tau_1, \tau_2, \cdots, \tau_\kappa$ at two month intervals each, assuming all subjects attend these check-ups. Following that, we check to see if the uncensored lifetimes, t_i falls in any of these intervals. If t_i falls in the interval (τ_m, τ_{m+1}) where $m \leq \kappa$, then the corresponding L_i and R_i the i^{th} observation will be τ_m and τ_{m+1}, respectively. Otherwise, if $t_i > \tau_\kappa$, t_i will be right censored at τ_κ.

Let $y_{L_i} = \log L_i$ and $y_{R_i} = \log R_i$ be the log left and log right endpoints for the i^{th} subject, where subject is right-censored at L_i if $L_i < t_i < \infty$. It follows that if $z_{L_i} = \frac{y_{L_i} - \boldsymbol{\beta' x_i}}{\sigma}$ and $z_{R_i} = \frac{y_{L_i} - \boldsymbol{\beta' x_i}}{\sigma}$, the log-likelihood function of the full sample with right- and interval-censored data for the LLR and EMVR distributions are as follows.

$$L(\boldsymbol{\beta}, \sigma) = \sum_{i=1}^{n} \delta_i \Bigg\{ \log\left[\frac{1}{1+e^{z_{L_i}}} - \frac{1}{1+e^{z_{R_i}}}\right] \Bigg\} + (1-\delta_i) \Bigg\{ \log\left[\frac{1}{1+e^{z_{L_i}}}\right] \Bigg\},$$

$$L(\boldsymbol{\beta}, \sigma) = \sum_{i=1}^{n} \delta_i \Big\{ \log\left[e^{-\exp(z_{L_i})} - e^{-\exp(z_{R_i})}\right] \Big\} - (1-\delta_i) \Big\{ \exp(z_{L_i}) \Big\}.$$

The parameter estimats can be obtained by solving the likelihood equations using any iterative procedure for solving non linear equations. In this research the maximum likelihood estimators of all the parameters were computed using the Newton Raphson iterative method.

The second simulation study focuses on the performance of the modified Cox-Snell residuals as the values of the martingale and deviance residual depends on these modified Cox-Snell residuals. The bootstrap simulation was carried using $B = 800$ for each replication. The purpose of this simulation study is to compare the performances of the modified Cox-Snell residuals based on the bias-corrected bootstrap harmonic mean and random imputation (Hboot) to the traditional Cox-Snell residual (CS) and modified Cox-Snell residual (MCS) proposed by Crowley and Hu (1977) using mid-point imputation.

To evaluate the performance of the different modifications of the Cox-Snell residuals, we need to obtain the estimated Kaplan-Meier survivor function based on the values of these modified residuals. Let $\hat{S}(r^*_{C_i})$ denote the estimated Kaplan-Meier survivor function based on the adjusted Cox-Snell residuals or Hboot. The plot of $\log[-\log(\hat{S}(r^*_{C_i}))]$ against $\log(r^*_{C_i})$ should be a straight line with unit slope and 0 intercept, if the residuals are performing well at indicating a good model fit. Thus, by implementing the same steps for the other residuals, their performances can be compared based on the mean absolute deviation (MAD) of the values of the intercept, slope, and coefficient of correlation, R from the desired value of 0, 1 and, 1 respectively.

Simulation results and conclusion

The results of the first simulation study are given in Tables 1-4. The tables display the bias and MSE of the estimates at different values of censoring proportions and sample sizes for both models. All values of bias and RMSE are relatively low, indicating good performance of the parameter estimates. The MSE of all parameter estimates generally increase with the increase in censoring proportions and decrease with the increase in sample sizes. As for the bias, it generally decreases with the increase in sample sizes but the trend when censoring proportion increases is not very clear. Some of the low values of bias at higher levels of censoring proportion do not imply that we have better estimates than the ones at lower censoring proportion. This is because the higher MSE values at higher censoring proportions suggest that the estimates are still typically far from the true value even though the average is close to the true parameter value.

Table 1: Bias and MSE for parameter estimates of LLR model when $n = 50$

	$\hat{\beta}_0$		$\hat{\beta}_1$		$\hat{\sigma}$	
cp	**Bias**	**MSE**	**Bias**	**MSE**	**Bias**	**MSE**
0.30	0.018735	0.022072	-0.032252	0.029920	-0.004607	0.007087
0.40	0.020070	0.024927	-0.027905	0.033175	-0.010090	0.007729
0.50	0.020132	0.029241	-0.023311	0.036426	-0.015340	0.008969
0.55	0.022437	0.034441	-0.021755	0.040738	-0.018476	0.010069
0.65	0.027842	0.044566	-0.018476	0.049655	-0.026861	0.013146

The results of the second simulation study are given in Figures 1-4. It is rather clear that the Hboot residuals consistently give much lower MAD values for intercept, slope and R at all levels of censoring proportions and sample sizes, for both LLR and EMVR models. The performance of the MCS is slightly better than CS residuals, however Hboot clearly outperforms both these residuals at indicating a good model fit.

Table 2: Bias and MSE for parameter estimates of LLR model when $n = 120$

	$\hat{\beta}_0$		$\hat{\beta}_1$		$\hat{\sigma}$	
cp	**Bias**	**MSE**	**Bias**	**MSE**	**Bias**	**MSE**
0.30	0.018141	0.009227	-0.018997	0.012033	0.005618	0.003015
0.40	0.017882	0.010286	-0.015034	0.013373	0.002225	0.003440
0.50	0.015741	0.011480	-0.007534	0.014333	-0.002945	0.003721
0.55	0.013322	0.012882	-0.003432	0.015260	-0.006945	0.004141
0.65	0.010740	0.016669	-0.006945	0.018069	-0.013634	0.005109

Table 3: Bias and MSE for parameter estimates of EMVR model when $n = 50$

	$\hat{\beta}_0$		$\hat{\beta}_1$		$\hat{\sigma}$	
cp	**Bias**	**MSE**	**Bias**	**MSE**	**Bias**	**MSE**
0.30	0.011842	0.034653	-0.007126	0.036929	-0.012768	0.024466
0.40	0.020372	0.042614	-0.003196	0.042867	-0.012853	0.026957
0.50	0.017397	0.052324	-0.000246	0.048838	-0.026407	0.032612
0.55	0.015044	0.060752	-0.002470	0.055881	-0.036834	0.035768
0.65	0.009622	0.094285	-0.036834	0.070551	-0.053256	0.046550

Table 4: Bias and MSE for parameter estimates of EMVR model when $n = 120$

	$\hat{\beta}_0$		$\hat{\beta}_1$		$\hat{\sigma}$	
cp	**Bias**	**MSE**	**Bias**	**MSE**	**Bias**	**MSE**
0.30	0.014489	0.013995	0.000465	0.015129	0.004680	0.010911
0.40	0.022461	0.017794	0.003017	0.017361	0.005203	0.011739
0.50	0.020638	0.021404	0.000426	0.020214	-0.005384	0.013203
0.55	0.016183	0.025140	0.000261	0.021883	-0.012736	0.014539
0.65	0.005314	0.030993	-0.012736	0.027365	-0.025983	0.016753

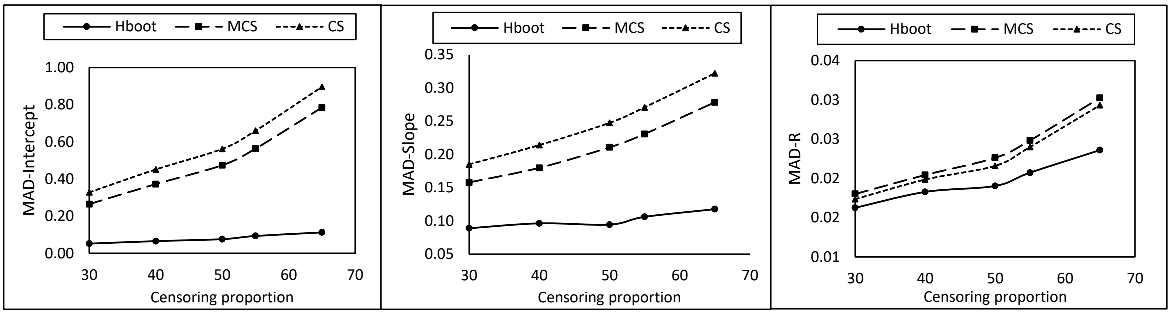


Figure 1: MAD for LLR model at $n = 50$

4. Real example of COM breast cancer data

The data on 94 breast cancer patients was discussed by Finkelstein and Wolfe (1985) where patients were given either radiation therapy (RT) or radiation therapy plus chemotherapy (RCT). The data is mixed case interval-censored as patients only have clinic visits that vary from patient to patient at different intervals. The interval-censored event time of the breast retraction is recorded by the physicians during the clinic visits. 38 patients who did not

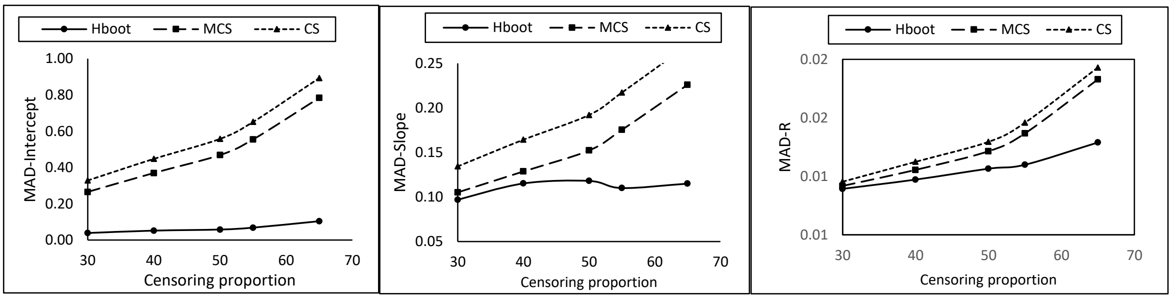


Figure 2: MAD for LLR model at $n = 120$

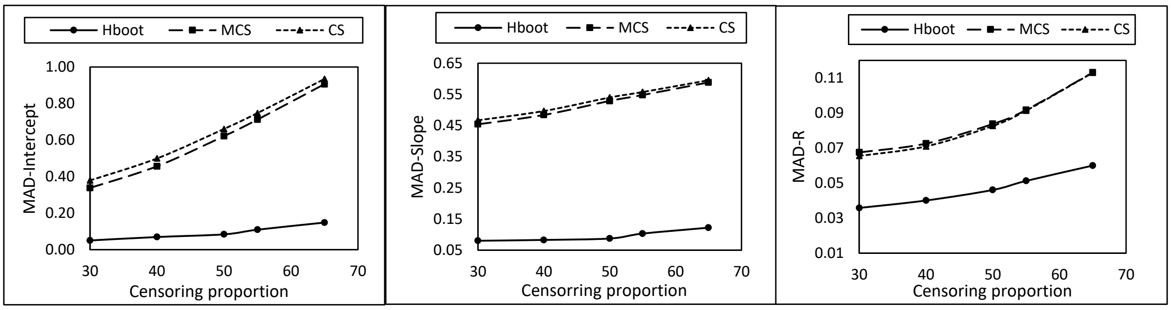


Figure 3: MAD for EMVR model at $n = 50$

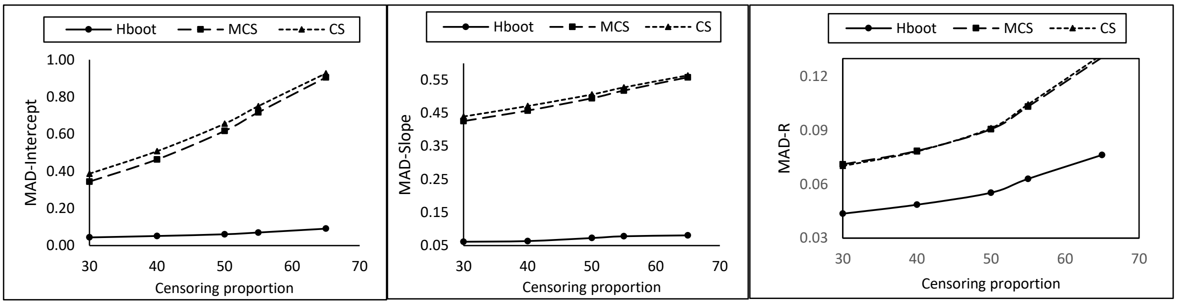


Figure 4: MAD for EMVR model at $n = 120$

experience breast retraction are right-censored observations.

To assess the model fit, we can obtain the nonparametric estimated survival probability proposed by Turnbull (1974) and compare it with the parametric survival probabilities obtained using the proposed model for different covariate values. We find that the EMVR model provides a good fit for the COM data. Figure 5 displays the plot of the estimated Turnbull survival probabilities overlapped with the survival probabilities obtained using the EMVR model for different covariate values. Both plots suggest that the use of the EMVR model will be appropriate for the data set. The survival functions for the two therapies indicate that the RT group has a longer time to breast retraction than the RCT group.

Table 5 shows the parameter estimates when the EMVR model was fit to the COM data with therapy as the covariate (RCT=1, RT=2). The p-value for β_1 suggests that there is a significant treatment effect. Figure 6 shows the estimated hazard plot (a) and survival probability plot (b) for the COM data using the imputed lifetimes. The hazard plot indicates that the RT group has a lower hazard (longer survival) compared to the RCT group. From the estimated parameters, the median survival time of the RT and RCT groups are 39.3 and 22.3 months respectively. The time ratio of the RT group to the RCT group at the median

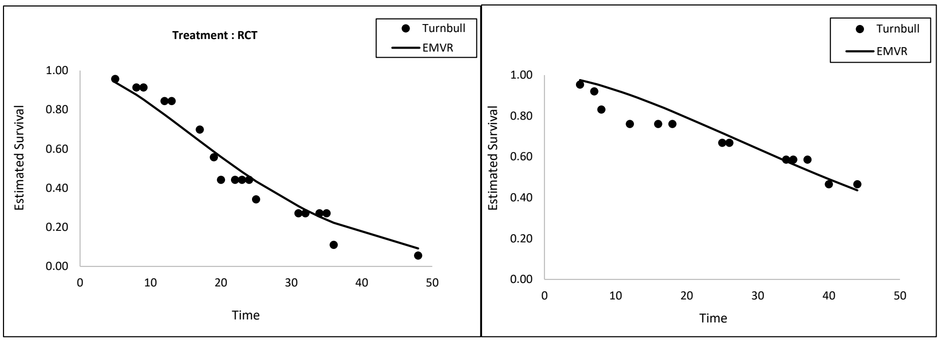


Figure 5: Turnbull and EMV Survivor Function Estimates for COM Data

Table 5: Estimates and 95% Wald interval for the parameters of EMVR Model

Parameter	Estimates	Std.Err	Z	P Val	lower	upper
β_0	2.76385	0.25424	10.871	0.000	2.265	3.262
β_1	0.56782	0.17568	3.232	0.001	0.223	0.9121
σ	0.61941	0.07427	8.340	0.000	0.473	0.765

survival time is 1.76. The EMVR model is a proportional hazards model and the hazard ratio of the RCT group to the RT group at any time is 2.50, indicating that the RCT group is failing at a rate more than twice that of the RT group.

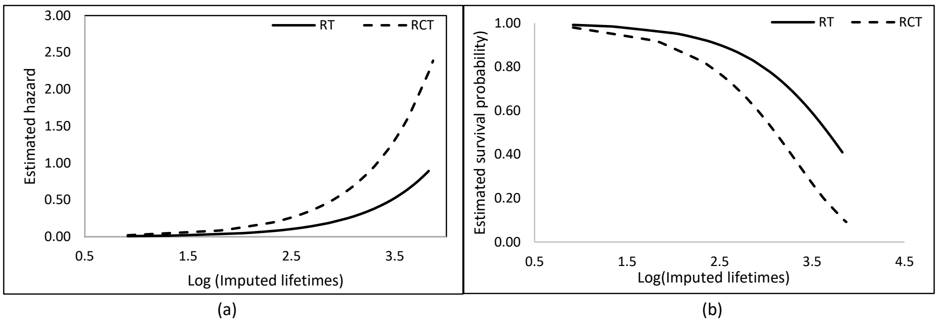


Figure 6: Estimated hazard (a) and survival probability (b) for COM Data

4.1. Adjusted residual plots

To check if an apt model has been fit to the data, we can use the plot of $\log[-\log(\hat{S}(r^*_{C_i}))]$ against $\log(r^*_{C_i})$, as shown in Figure 7. The estimated intercept and slope values are -0.1234 and 0.9130 respectively, which are very close to the ideal values of 0 and 1, indicating that the EMVR model fits the data very well.

The plot of the adjusted martingale residuals versus observation number (a) and versus treatment type (b) are shown in Figure 8. Both plots indicate that there might be an outlier which corresponds to patient 94 whose failure times is greater than 48 months, which is much larger than other patients who received RCT treatment.

The plot of the adjusted deviance residuals versus observation number (a) and versus treatment type (b) are shown in Figure 9. The outlying effect of observation 94 is not that apparent in both these plots. This is because the deviance residuals are more symmetrically distributed around 0 if an apt model has been fit to the data. The score residual for the EMVR model

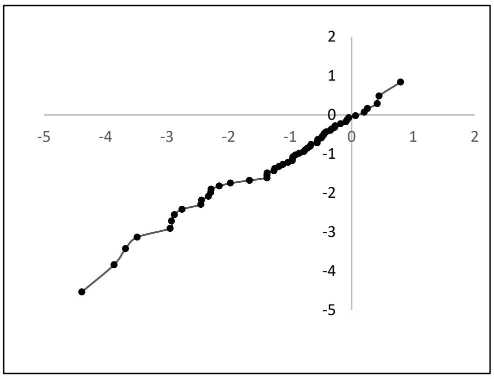


Figure 7: Index plot of adjusted Cox-Snell residuals

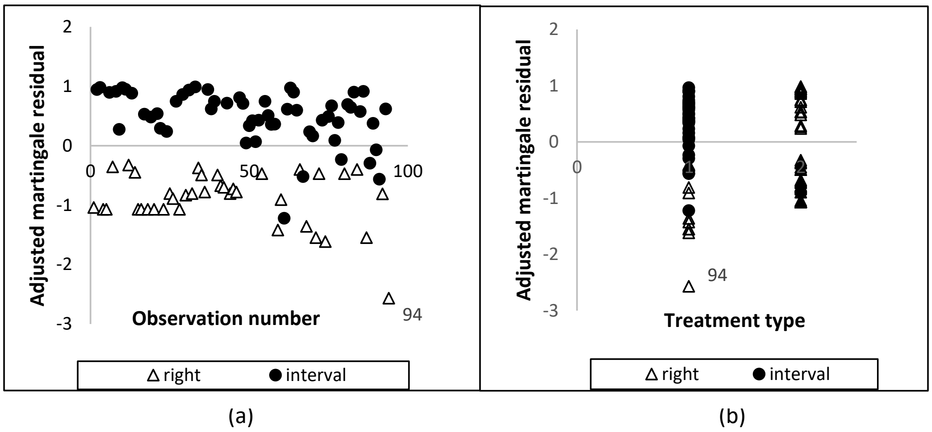


Figure 8: Adjusted martingale residuals vs observation number, (a) and vs treatment, (b)

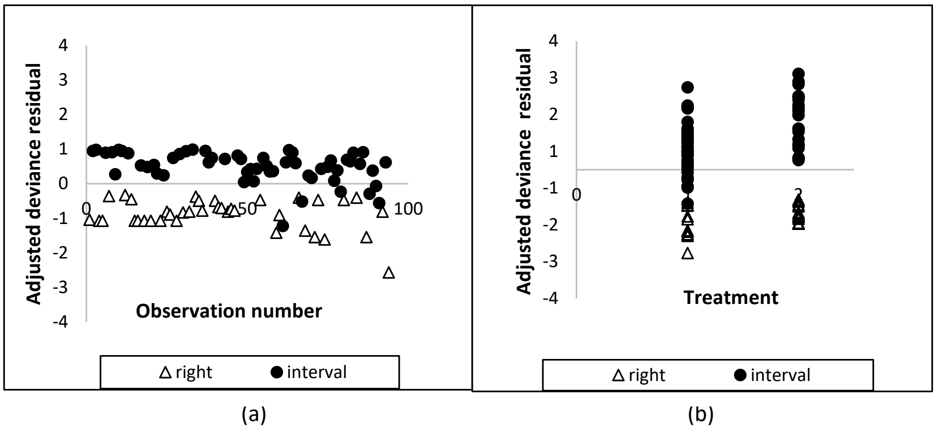


Figure 9: Adjusted deviance residuals vs observation number (a) and vs treatment (b)

with right and interval-censored data are the components of the first derivatives of the log-likelihood function with respect to its parameters, β and σ evaluated at their respective MLEs as given in the following,

$$\frac{\partial L(\beta,\sigma)}{\partial \beta_j} = \sum_{i=1}^{n} \frac{x_{ij}}{\sigma}\left[\frac{\delta_i\left(e^{z_{L_i}-\exp(z_{L_i})} - e^{z_{R_i}-\exp(z_{R_i})}\right)}{e^{-\exp(z_{L_i})} - e^{-\exp(z_{R_i})}} + (1-\delta_i)e^{z_{L_i}})\right],$$
$$j = 0, 1\cdots, p-1,$$

$$\frac{\partial L(\beta,\sigma)}{\partial \sigma} = \sum_{i=1}^{n} \frac{1}{\sigma}\left[\frac{\delta_i\left(z_{L_i}e^{z_{L_i}-\exp(z_{L_i})} - z_{R_i}e^{z_{R_i}-\exp(z_{R_i})}\right)}{e^{-\exp(z_{L_i})} - e^{-\exp(z_{R_i})}} + (1-\delta_i)z_{L_i}e^{z_{L_i}})\right].$$

Another simpler technique to compute the adjusted score residuals for the EMVR model with right and interval-censored data is by using the imputed lifetimes discussed in Section 2.2. Let,

$$y_i^I = \begin{cases} \log(t_i^*) & \text{for } t_i \text{ interval-censored,} \\ \log(t_{L_i}) & \text{for } t_i \text{ right-censored.} \end{cases}$$

If $z_i^* = \frac{y_i^I - \boldsymbol{\beta}'\boldsymbol{x}_i}{\sigma}$, the adjusted score residuals ($r_{S_i}^*$) can now be calculated from the following components of the first derivatives of the log-likelihood function with respect to its parameters, β and σ evaluated at their respective MLEs.

$$\frac{\partial L(\beta,\sigma)}{\partial \beta_j} = \sum_{i=1}^{n} \frac{x_{ij}}{\sigma}\left(e^{z_i^*} - \delta_i\right), j = 0, 1\cdots, p-1,$$

$$\frac{\partial L(\beta,\sigma)}{\partial \sigma} = \sum_{i=1}^{n} \frac{z_i^* e^{z_i^*}}{\sigma} - \delta_i\left(\frac{z_i^*+1}{\sigma}\right).$$

The plot of the score residuals based on imputed lifetimes versus observation number (a) and versus treatment type (b) for the covariate treatment, are shown in Figure 10. We do not observe any pattern that raises concern except that the score residual value for observation 94 is slightly larger than other observations.

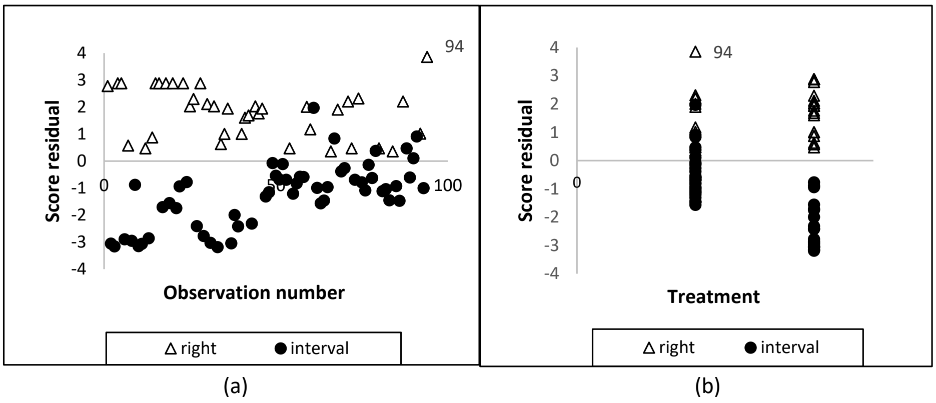


Figure 10: Score residuals vs observation number (a) and vs treatment (b) using imputed lifetimes

The case-weights perturbation scheme for the EMVR model would yield the following log-likelihood function,

$$L(\boldsymbol{\beta},\sigma) \;=\; \sum_{i=1}^{n}\omega_i\delta_i\left\{\log\left[e^{-\exp(z_{L_i})}-e^{-\exp(z_{R_i})}\right]\right\}-\omega_i(1-\delta_i)\left\{\exp(z_{L_i})\right\}.$$

Following if $\hat{\boldsymbol{\theta}}$ and $\hat{\theta}_{(\omega)}$ are the MLEs under $L(\boldsymbol{\theta})$ and $L(\boldsymbol{\theta}|\boldsymbol{\omega})$, the likelihood displacement statistic to assess the influence of $\boldsymbol{\omega}$, is given by the following,

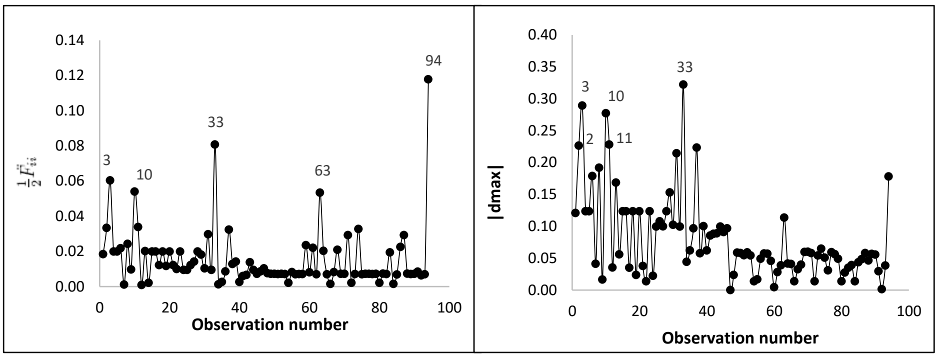
$$LD(\boldsymbol{\omega}) = 2[\log L(\hat{\boldsymbol{\theta}}) - \log L(\hat{\boldsymbol{\theta}}_{(\omega)})].$$

and the $(k,i)^{\text{th}}$ element of Δ_{ki} given by the following,

$$\Delta_{ki} = \begin{cases} \frac{x_{ij}}{\sigma}\left[\frac{\delta_i\left(e^{z_{L_i}-\exp(z_{L_i})}-e^{z_{R_i}-\exp(z_{R_i})}\right)}{e^{-\exp(z_{L_i})}-e^{-\exp(z_{R_i})}}+(1-\delta_i)e^{z_{L_i}}\right], \\ k=1,2\cdots,p,\;\; i=1,2,\ldots,n,\;\; j=k-1, \\ \\ \frac{1}{\sigma}\left[\frac{\delta_i\left(z_{L_i}e^{z_{L_i}-\exp(z_{L_i})}-z_{R_i}e^{z_{R_i}-\exp(z_{R_i})}\right)}{e^{-\exp(z_{L_i})}-e^{-\exp(z_{R_i})}}+(1-\delta_i)z_{L_i}e^{z_{L_i}}\right], \\ k=p+1,\;\; i=1,2,\ldots,n. \end{cases}$$

Again, it would be much simpler to use y_i^I, the imputed lifetime for ease of computation when data is interval-censored. If $z_i^*{=}\frac{y_i^I-(\beta_0+\beta_1 x_{1i})}{\sigma}$, we could use the modified likelihood based on the imputed lifetime to obtain the $(k,i)^{\text{th}}$ element of Δ_{ki}^* given by the following,

$$\Delta_{ki}^* = \begin{cases} \frac{x_{ij}}{\sigma}\left(e^{z_i^*}-\delta_i\right),\;\; k=1,2\cdots,p,\;\; i=1,2,\ldots,n,\;\; j=k-1, \\ \\ \frac{z_i^* e^{z_i^*}}{\sigma}-\delta_i\left(\frac{z_i^*+1}{\sigma}\right),\;\; k=p+1,\;\; i=1,2,\ldots,n. \end{cases}$$



(a) (b)

Figure 11: Index plot of $\frac{1}{2}\ddot{F}_{ii}$ (a) and $|d_{max}|$ (b), using imputed lifetime

Figure 11 shows the index plot of $\frac{1}{2}\ddot{F}_{ii}$, which can be used to assess the influence of the total parameter vector $\boldsymbol{\theta}$. The figure indicates that observations 3,10,33,63 and 94 may be potentially more influential than other observations. Although none of these observations exceed the warning limit of $\frac{1}{2}\chi^2_{(0.5;p+1)} = 1.183$, it's interesting to know why these observations especially 3,10 and 33 are more influential. Observation 94, is also the only observation that was singled out by the martingale, deviance, and score residual plots. This patient has the largest right-censored lifetime at 48 months which is much longer compared to other patients who received RCT treatment. The failure time of patient 33 is lesser than 5 months, which is the minimum among patients receiving RT treatment. Patients 3 and 10 also received RT treatment, and have failure times lesser than 7 and 8 months, respectively. Although observation 63 has failure time lesser than 5 months, it was not as influential as patient 33 because the patient received RCT treatment.

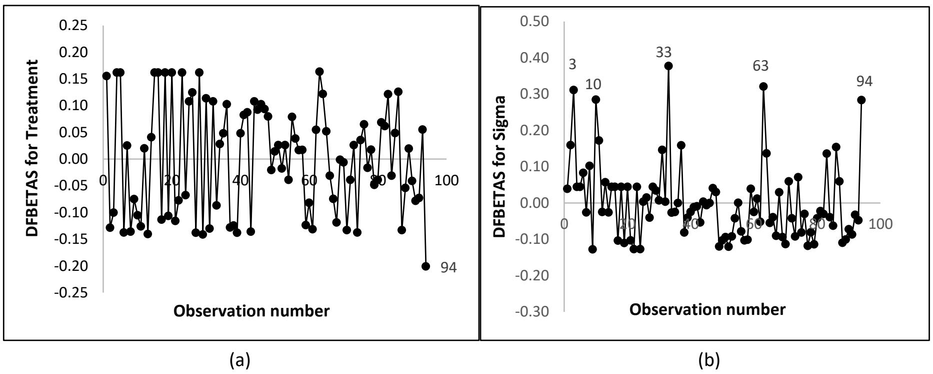


Figure 12: DFBETAS for treatment vs observation number

As discussed earlier, the index plot of $\frac{1}{2}\ddot{F}_{ii}$ gives an approximation to the information given by the DFBETAS. To check this, we can observe the index plot of the DFBETAS for the estimated coefficient of the covariate treatment and σ, shown in Figure 4.1. We can see that the single case perturbation does not seem to have any influential effect on the estimated coefficient of the covariate treatment, except for observation 94 which has a value that is larger than 0.20 which is very close to the usual cutoff value of $2/\sqrt{94} = \pm 0.206$. However, the index plot of the DFBETAS for the parameter σ indicates all the observations discussed earlier, namely 3,10,33,63 and 94 as being potentially influential. Thus, what is obvious is that the single case perturbation mainly affects the estimated parameter σ, but not the other parameters.

The index plot of d_{max} shown in Figure 11 identifies all the point identified by the index plot of $\frac{1}{2}\ddot{F}_{ii}$, but it doesn't identify observation 63 and 94 as being highly influential. Instead, it identifies observations 2,3,10,11, and 33 as being the five most influential observations. If we inspect observations 2 and 11, they both received RT treatment and have failure times lesser than 4 and 6 months, respectively. However there are several other observations with values that are rather similar, hence we can conclude that there is nothing highly unusual about these two patients. This clearly shows that the $\frac{1}{2}\ddot{F}_{ii}$ plot is better at detecting the effects of single case perturbations compared to the d_{max} plot.

Next, it would be of interest to know the impact of removing groups of these influential observations on the estimated parameters, median time ratios, and hazard ratios. Observation 94 is the only case singled out by all the residual and influence diagnostics plots. When this observation is removed, the median survival time of the RT and RCT groups change to 41.9 and 22.6 months respectively, and the median time ratio of the RT group to the RCT group increases to 1.85. The estimated hazard ratio of the RCT group to the RT group changes from 2.50 to 2.74.

When the most influential observations detected by the $\frac{1}{2}\ddot{F}_{ii}$ and d_{max} plots, namely observations 94 and 33 are deleted, the median survival time of the RT and RCT groups change to 40.1 and 21.7 months respectively, and the median time ratio of the RT group to the RCT group stays at 1.85. The estimated hazard ratio of the RCT group to the RT group changes from 2.50 to 2.95, which is rather substantial. It implies that the RCT group is failing at a rate almost three times that of the RT group. Although it does not change the overall conclusion that the RT treatment works better than RCT treatment, it may provide stronger evidence that the RT treatment alone works far better in delaying the onset of breast retraction for breast cancer patients.

It would not be wise to simply remove other observations because they are identified as being influential as they still carry valuable information regarding the treatment administered. For example, observations 3,10 and 33 show us that some patients can still have breast retraction as early as 5-8 months, even while receiving the RT treatment, although the median survival time is 39.3 months. Plots of the score residuals, $\frac{1}{2}\ddot{F}_{ii}$ and d_{max} using interval censored lifetimes are shown in Appendix A. These plots look almost identical to the ones obtained using imputed lifetimes and provide us with the same information but require significantly more time and effort to produce.

5. Conclusion

In this research, we explored the residuals based on the bias-corrected bootstrap estimates and a random imputation technique to deal with right and interval-censored lifetime data. The proposed residual works well even when data has mixed case censoring and can be easily applied to other models because it's a computer intensive technique that is free from any theoretical assumptions. These estimates are obtained directly from the data and are better estimates of the excess, ζ, in the Cox-Snell residuals, compared to the traditional and modified Cox-Snell residuals as indicated by the simulation results.

The proposed randomly imputed lifetimes are very useful when we are dealing with interval-censored data, where the derivation of exact likelihood contributions can become rather tedious and unnecessarily lengthy for some models such as models involving time dependent covariates, truncated data and bivariate models with right- and interval-censored data. These imputed lifetimes are simple and easy to implement due to the recent advancement in computing technology and are very useful and practical in assessing model adequacy and identify potential outliers and influential observations. The local influence diagnostics using these imputed lifetimes are easy to implement and requires far lesser time and effort.

Acknowledgement

We would like to thank Universiti Putra Malaysia for their financial support of this research work via the PUTRA GRANT, VOT 9595300 .

References

Abramovitch L, Singh K (1985). “Edgeworth Corrected Pivotal Statistics and the Bootstrap.” *Annals of Statistic*, **13**, 116–132. URL https://doi.org/10.1214/aos/1176346580.

Arasan J, Lunn M (2008). “Alternative Interval Estimation for Parameters of Bivariate Exponential Model with Time Varying Covariate.” *Computational Statistics*, **23**, 605–622. URL https://doi.org/10.1007/s00180-007-0101-9.

Arasan J, Lunn M (2009). “Survival Model of A Parallel System With Dependent Failures And Time Varying Covariates.” *Journal of Statistical Planning and Inference*, **139**(3), 944–951. URL https://doi.org/10.1016/j.jspi.2008.06.007.

Barlow WE, Prentice RL (1988). “Residuals for Relative Risk Regression.” *Biometrika*, **79**(1), 65–74. URL https://doi.org/10.1093/biomet/75.1.65.

Collett D (2015). *Modelling Survival Data in Medical Research*. CRC press. ISBN 978-1584883258, URL https://www.amazon.com/Modelling-Survival-Medical-Research-Second/dp/1584883251.

Cook RD (1986). "Assessment of Local Influence." *Journal of the Royal Statistical Society: Series B (Methodological)*, **48**(2), 133–155. URL https://www.jstor.org/stable/2345711.

Crowley J, Hu M (1977). "Covariance Analysis of Heart Transplant Survival Data." *Journal of the American Statistical Association*, **72**(357), 27–36. URL https://doi.org/10.2307/2286902.

Efron B, Tibshirani RJ (1994). *An Introduction to the Bootstrap*. CRC press. ISBN 978-0412042317, URL https://www.routledge.com/An-Introduction-to-the-Bootstrap/Efron-Tibshirani/p/book/9780412042317.

Escobar LA, Meeker Jr WQ (1992). "Assessing Influence in Regression Analysis with Censored Data." *Biometrics*, pp. 507–528. URL https://doi.org/10.2307/2532306.

Farrington C (2000). "Residuals for Proportional Hazards Models with Interval-Censored Survival Data." *Biometrics*, **56**(2), 473–482. URL https://onlinelibrary.wiley.com/doi/10.1111/j.0006-341X.2000.00473.x.

Finkelstein DM, Wolfe RA (1985). "A Semiparametric Model for Regression Analysis of Interval-Censored Failure Time Data." *Biometrics*, pp. 933–945. URL https://doi.org/10.2307/2530965.

Grzegorzewski P, Hryniewicz O, Romaniuk M (2020). "Flexible Resampling for Fuzzy Data." *International Journal of Applied Mathematics and Computer Science*, **30**(2). URL https://www.researchgate.net/publication/342657728_Flexible_resampling_for_fuzzy_data.

Hosmer DW, Lemeshow S, May S (2008). *Applied Survival Analysis: Regression Modelling of Time-to-Event Data*. Wiley-Interscience. 978-0471754992, URL https://www.wiley.com/en-us/Applied+Survival+Analysis%3A+Regression+Modeling+of+Time+to+Event+Data%2C+2nd+Edition-p-9780471754992.

Kalbfleisch JD, Prentice RL (2011). *The Statistical Analysis of Failure Time Data*. John Wiley & Sons. ISBN 978-1118032985, URL https://onlinelibrary.wiley.com/doi/book/10.1002/9781118032985.

Lagakos SW (1981). "The Graphical Evaluation of Explanatory Variables in Proportional Hazard Regression Models." *Biometrika*, **68**(1), 93–98. URL https://doi.org/10.1093/biomet/68.1.93.

Lai MC, Arasan J (2020). "Single Covariate Log-Logistic Model Adequacy with Right And Interval Censored Data." *Journal of Quality Measurement and Analysis*, **16**(2), 131–140. URL http://journalarticle.ukm.my/16065/1/jqma-16-2-paper1.pdf.

Law CG, Brookmeyer R (1992). "Effects Of Mid-Point Imputation on The Analysis Of Doubly Censored Data." *Statistics in medicine*, **11**(12), 1569–1578. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.4780111204.

Lesaffre E, Verbeke G (1998). "Local Influence in Linear Mixed Models." *Biometrics*, pp. 570–582. URL https://doi.org/10.2307/3109764.

Lui KJ, Darrow WW, Rutherford III GW (1988). "A Model-Based Estimate of The Mean Incubation Period for Aids in Homosexual Men." *Science*, **240**(4857), 1333–1335. URL https://pubmed.ncbi.nlm.nih.gov/3163848.

Manoharan T, Arasan J, Midi H, Adam MB (2017). "Bootstrap Intervals in the Presence of Left-Truncation, Censoring and Covariates with A Parametric distribution." *Sains Malaysiana*, **46**(12), 2529–2539. URL http://www.ukm.my/jsm/pdf_files/SM-PDF-46-12-2017/31%20Thirunanthini.pdf.

Manoharan T, Arasan J, Midi H, Adam MB (2020). "Influential Measures on Log-Normal Model for Left-Truncated and Case-K Interval Censored Data With Time-Dependent Covariate." *Communications in Statistics-Simulation and Computation*, **49**(6), 1445–1466. URL https://doi.org/10.1080/03610918.2018.1498885.

Mariotto AB, Mariotti S, Pezzotti P, Rezza G, Verdecchia A (1992). "Estimation of the Acquired Immunodeficiency Syndrome Incubation Period in Intravenous Drug Users: A Comparison with Male Homosexuals." *American Journal of Epidemiology*, **135**(4), 428–437. URL https://doi.org/10.1093/oxfordjournals.aje.a116303.

Meeker WQ, Escobar LA, Pascual FG (2014). *Statistical Methods for Reliability Data, 2nd Edition*. John Wiley & Sons. ISBN: 978-1118115459, URL https://www.wiley.com/en-us/Statistical+Methods+for+Reliability+Data%2C+2nd+Edition-p-9781118115459.

Naslina AMNN, Jayanthi A, Syahida ZH, Bakri AM (2020). "Assessing the Goodness of Fit of the Gompertz Model in the Presence of Right and Interval Censored Data with Covariate." *Austrian Journal of Statistics*, **49**(3), 57–71. URL https://doi.org/10.17713/ajs.v49i3.1085.

Ortega EM, Cancho VG, Bolfarine H (2006). "Influence Diagnostics in Exponentiated-Weibull Regression Models with Censored Data." *SORT-Statistics and Operations Research Transactions*, pp. 171–192. URL https://www.researchgate.net/publication/28175480_Influence_diagnostics_in_exponentiated-Weibull_regression_models_with_censored_data.

Pettitt A, Daud IB (1989). "Case-Weighted Measures of Influence for Proportional Hazards Regression." *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **38**(1), 51–67. URL https://doi.org/10.2307/2347680.

Schoenfeld D (1982). "Partial Residuals for The Proportional Hazards Regression Model." *Biometrika*, **69**(1), 239–241. URL https://www.wiley.com/en-us/Applied+Survival+Analysis%3A+Regression+Modeling+of+Time+to+Event+Data%2C+2nd+Edition-p-9780471754992.

Singh K (1981). "On the asymptotic accuracy of Efron's bootstrap." *The Annals of Statistics*, **9**(6), 1187–1195. URL https://doi.org/10.1214/aos/1176345636.

Sun J (2006). *The statistical analysis of interval-censored failure time data*, volume 3. Springer. 978-0387371191, URL https://www.springer.com/gp/book/9780387329055.

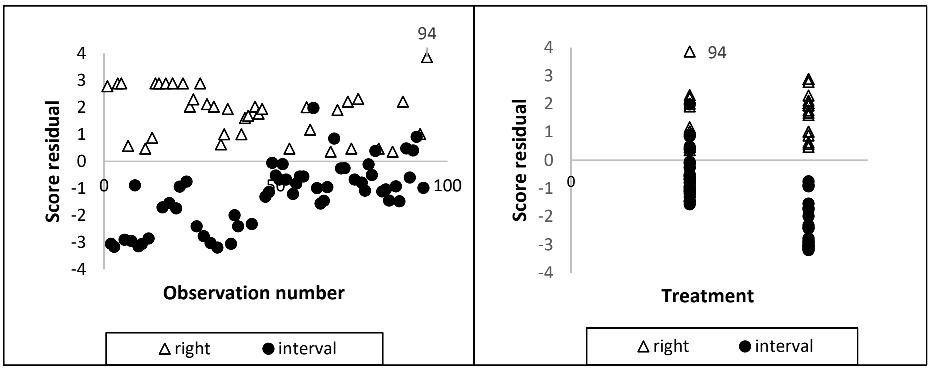
Therneau TM, Grambsch PM, Fleming TR (1990). "Martingale-Based Residuals for Survival Models." *Biometrika*, **77**(1), 147–160. URL https://doi.org/10.2307/2336057.

Turnbull BW (1974). "Nonparametric Estimation of a Survivorship Function with Doubly Censored Data." *Journal of the American statistical association*, **69**(345), 169–173. URL https://doi.org/10.2307/2285518.

Weissfeld LA, Schneider H (1990). "Influence Diagnostics for the Weibull Model Fit to Censored Data." *Statistics & probability letters*, **9**(1), 67–73. URL https://deepblue.lib.umich.edu/bitstream/handle/2027.42/28787/0000621.pdf?sequence=1.

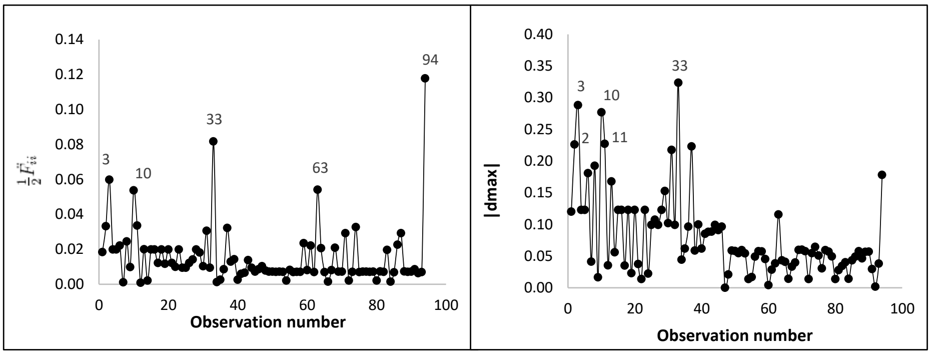
Zhuang L, Xu A, Pang J (2021). "Product Reliability Analysis Based on Heavily Censored Interval Data with Batch Effects." *Reliability Engineering & System Safety*, **212**, 107622. URL https://doi.org/10.1016/j.ress.2021.107622.

Appendix A



(a) (b)

Figure 13: Score residuals vs observation number (a) and vs treatment (b)



(a) (b)

Figure 14: Index plot of $\frac{1}{2}\ddot{F}_{ii}$ (a) and $|d_{max}|$ (b)

Pseudo code for the simulation study

This program produces the outputs of the simulation study with right and interval censored data.

1. Clean workspace using rm(list=ls()).

2. Call out the maxLik package using library(maxLik).

3. Set the initial values for sample size (n), parameter values (β_0, β_1, σ), and censoring proportion, CP (set μ).

4. Begin the loop for $N = 1000$ replication using for (i in 1:1000).

   - Simulate uniform variates from U(0,1) using u<-runif(n,0,1).
   - Generate the lifetime, t, using the inverse transform technique.
   - Simulate right censoring time, c, from exp(μ) - adjust μ to obtain desired approximate CP.
   - Determine the value of indicator variable, s,

     If ($t < c$) **then return** $s = 1$ **else** $s = 0$.

- Set lifetime as censoring time for censored observations, t=min(t, c).
- Convert uncensored observations to interval censored observations.
  - Compare each interval censored lifetimes with predefined intervals. For interval $(0, 2)$,

    If $(0 < t < 2)$, the left endpoint $L_i = 0$, right end point, $R_i = 2$.
  - Repeat for all the intervals until appropriate left and right endpoints are assigned to all uncensored data.
  - Observations greater than final inspection time, τ_κ will be censored.

    If $(t > \tau_\kappa)$, let $L_i = \tau_\kappa$ and change $s = 1$ to $s = 0$.
- Calculate the right censoring proportion, CP.
- Use function maxLik() to obtain the estimated parameters of (β_0, β_1, σ)
- **Implementation of random imputation for interval censored data.**
- Begin the loop for n replication for each observation.
  - Begin another loop for 1000 replication.
    - * Obtain $\hat{S}(R_i)$ and $\hat{S}(L_i)$ using the estimated parameters for the i^{th} interval censored observation.
    - * Simulate uniform variates from $U(\hat{S}(R_i), \hat{S}(L_i))$.
    - * Transform the uniform variate to obtain the estimated lifetime, t_i^r.
  - End of loop for 1000 replication.
  - Obtain t_i^*, the mean of t_i^r.
- End of loop for n replication
- Calculate the adjusted Cox-Snell residual (Hboot).
  - Create a new data frame consisting of the random imputed lifetimes and right censored lifetimes.
  - Generate 800 bootstrap samples of size n each from the adjusted Cox-Snell residuals.
  - Calculate the harmonic mean of each of the bootstrap sample.
  - Obtain the mean of the harmonic means of the bootstrap samples.
  - Calculate bias-corrected bootstrap harmonic mean, h_{bc}.
  - Adjust the Cox-Snell residual for right censored data, $r_{C_i}^{\text{R}} = r_{C_i} + h_{bc}$.
  - Combine Cox-Snell residual for right and interval censored data and call it Hboot.
- Calculate the traditional Cox-Snell residual (CS) and modified Cox-Snell residual (MCS).
  - Create a new data frame consisting of midpoint imputed lifetime (t_m) and right censored lifetimes.
  - Obtain traditional Cox-Snell residual (CS) for right censored data, $r_{C_i}^{\text{CS}} = r_{C_i} + 1$.
  - Obtain modified Cox-Snell residual (MCS) for right censored data, $r_{C_i}^{\text{CS}} = r_{C_i} + 0.693$.
  - Combine traditional and modified Cox-Snell residual for right and interval censored data and call it CS and MCS.
- Obtain the Kaplan-Meier estimates for Hboot, CS and MCS using survfit().
- Fit a regression line for $y = \log[-\log(\hat{S}(Hboot))]$ against $x = \log(Hboot)$ using lm() function.
- Fit a regression line for $y = \log[-\log(\hat{S}(CS))]$ against $x = \log(CS)$ using lm() function.

- Fit a regression line for $y = \log[-\log(\hat{S}(MCS))]$ against $x = \log(MCS)$ using lm() function.
- Obtain the coefficient for the slope, intercept, and correlation for all the fitted lines.

5. End of the loop for $N = 1000$ replication.
6. Calculate the average of CP.
7. Calculate the bias and MSE of the parameter estimates.
8. Calculate the MAD for slope, intercept, and correlation for all the residuals.

Affiliation:

Jayanthi Arasan
Department of Mathematics
Faculty of Science, UPM
Serdang, Selangor, Malaysia
E-mail: jayanthi@upm.edu.my

Austrian Journal of Statistics http://www.ajs.or.at/
published by the Austrian Society of Statistics http://www.osg.or.at/
Volume 52 *Submitted:* 2021-07-28
January 2023 *Accepted:* 2021-11-22