MSI Journal of
Multidisciplinary Research (MSIJMR)

ROBUST PROCEDURE FOR HANDLING CENSORED DATA IN CLINICAL TRIALS

Eric Boahen[1*]

[1*] Department of Statistics School of Mathematical Sciences University of Technology and Applied Sciences, Ghana.

*** Correspondence:** Eric Boahen

The authors declare that no funding was received for this work.

OPEN ACCESS

Received: 26-December-2025
Accepted: 30-January-2026
Published: 04-February-2026

This article is published in the **MSI Journal of Multidisciplinary Research (MSIJMR) ISSN 3049-0669 (Online)**

The journal is managed and published by MSI Publishers.

Volume: 3, Issue: 2 (February-2026)

ABSTRACT: One of the difficult aspect in parameter estimation in survival analysis is the presence of censored values in survival data When patients survival time are measured in continuous time interval, the censored values continue to create discrepancies in estimations since the stochastics realization of censored values are masked. For this reason, appropriate distributions must be specified before maximum likelihood is used. An optimal approach that merges both censored values and uncensored values is appreciable since asymptotically distributions are not normal. Robust model and efficient algorithms developed to enhance optimal performance in estimation. Simulations show that the optimal robust model is maximized.

Introduction

Survival analysis is generally defined as a set of methods for analyzing data where the outcome variable is the time until the occurrence of an event of interest. The event can be death, occurrence of a disease, marriage, divorce, etc. The time to event or survival time can be measured in days, weeks, years, etc. If the event of interest is heart attack, then the survival time can be the time in years until a person develops a heart attack.

In survival analysis, subjects are usually followed over a specified time period and the focus is on the time at which the event of interest occurs. Survival times are typically positive numbers; this means that ordinary linear regression may not be the best choice unless these times are first transformed in a way that removes this restriction. Second, and more importantly, ordinary linear regression cannot effectively handle the censoring of observations. Observations are called censored when the information about their survival time is incomplete; the most commonly encountered form is right censoring. A right censoring occurs when time to an event is greater than threshold value. A left censoring is when time to a particular event is less the threshold value or censoring limit. In interval censoring, observations are made at infrequent intervals and the time information of the timing events is not easily captured. Interval events are continuous in nature that needs specific procedure to handle the data. Interval estimations need to be tackled by resorting to discrete approach. Estimations and time monitoring are sometimes missing in the process of acquiring data. For this reason, survival data are mostly censored and part of information is incomplete. Estimating mean, variance and standard deviation from censored data is different from those in uncensored data. There is a margin of estimation bias because of incomplete information. Incomplete information is generated by several factors such as incomplete follow-up and staggered entry of patients into the study process. In the process of the study, some of the entities needed to be measured may still not happen at the time of analysis. For this reason, these entities are right censored (Allison,1995). In addition to this, following a particular event can be lost because of insufficient follow-up. Censored observations occur because inadequate time to follow-up; entities to be followed may either drop out of the study completely or missing from the study center or may move away to a different vicinity it cannot be located. Several factors competing for the occurrence of an event are also major contributing factors that generate censored observations in survival studies.

Introduction

The result of incomplete information in censored data is an estimation problem. An increase in the proportion of censored observations in an entire survival data equally

increases the margin of bias in statistical estimates . As bias increases, precision decreases and this eventually affect validity of statistical estimates. Affected results in survival estimates have an effect on statistical power leading to questionable reliability of statistical inferences. In survival analysis, already generated censored data are affected by estimation bias. This bias cannot be perfectly avoided but can be minimized. For this reason, statisticians have a duty to develop best models so as to reduce the bias. Estimating both left and right censored data call for general method which results in interval estimations. The continuous nature of estimating parameters in intervals is cumbersome and frustrating. For this reason, an algorithm that can convert continuous event time to discrete is preferred. In censored data analyses, the observed samples used are very few, this is because clinical data are mostly censored and part of it is incomplete. In this situation the exponential Distribution is very useful. Some events in survival analysis, especially in censoring studies just occur in natural order, as a result of this, cost and time are reduced. If first, second and third order of failures are recorded; there is no need to wait for all the samples to be covered. The few initial occurrences can be used in exponential distribution to enhance maximum information in estimation. We therefore propose exponential distribution to estimate patients' survival time. Two points account for the length of time a patient spends in the study; these are the starting point and the time of death. Occasionally, the time of death is obvious but the choice of the proper starting time is not clear because the starting time for admitting patients into the study depends on the time they report for treatment. To estimate parameters in a general problem like this, there is the need to start all patients at a common point.

Robustness of the general model

In survival analysis, patients' staggered entry point poses estimation difficulties and analysis is complicated in censored data. The proposed general model in this work accounted for the individual patients staggered starting point by starting all patients at a common point, usually time zero. This makes the model efficient, an unbiased estimator, have minimum variance and robust. A robust procedure is a procedure where the accuracy of the procedure does not depend too heavily on the distribution assumptions being true. This means that the general model developed here can

accommodate all kinds of data. We propose a general model of $f(\pi_i) = \lambda^r e^{-\lambda V}$, where v is the sum of patients time in recursive intervals for both observed and censored values, λ is the patients hazard rate, π_i is the total number of patients observed in the i^{th} recursive interval, r is the total number of patients who die in the course of the study.

$\log f(\pi) = r log\lambda - \lambda V + C$

$\frac{dL}{d\lambda} = \frac{r}{\lambda} - V$. At $\frac{dL}{d\lambda} = 0, \hat{\lambda} = \frac{r}{V}$

$\hat{\lambda} = \frac{r}{v}$ is also a function of v alone.To ascertain whether $\hat{\lambda}$ is also minimum variance unbiased estimator, the expectation of $\hat{\lambda}$ is estimated under the procedure of logarithm of $\hat{\lambda}$. This is because, $\hat{\lambda}$ has a normal distribution with mean $\log(\hat{\lambda})$ and variance $\frac{\sigma^2}{r}$, the data follow $f(\pi_i) = \lambda^r e^{-\lambda V}$ the exponential distribution, therefore $X^2 = 2\lambda v = 2r\lambda\frac{v}{r} = 2r\,{}^{\lambda}/_{\hat{\lambda}}$. In terms of $\hat{\lambda}$, $\hat{\lambda} = \frac{2r\lambda}{X^2}$, where X^2 has a chi-square distribution with 2r degrees of freedom. This is because most text tabulates the chi-square distribution but not the gamma distribution. In this situation, a X^2 table is useful to find probability for gamma variates. If V follows a gamma distribution with parameters λ and r, then, it follows that

$P[V \leq v] = P[X^2 \leq 2\lambda V]$.

$\log(\hat{\lambda}) = log2r + \log\lambda - \log X^2$.

The moment of $\log \hat{\lambda}$ becomes useful because the mean and variance of $\log \hat{\lambda}$ largely depend on the moment of $\log X^2$. In the theory of moment statistics, a variant X^2 with chi-square distribution has v degrees of freedom, with the expected value of the variant being $E(X^2) = v$. The variance $v(X^2) = 2v$. $E(X^2 - v)^3 = 8v$ and $E(X^2 - v)^4 = 48v + 12v^2$. The bases of this is because, the chi-square distribution is a special case of the gamma distribution, where $\alpha = \frac{v}{2}, \beta = 2$. In relating chi-square to standardized normal distribution

$\emptyset = \frac{(X^2 - v)}{\sqrt{2v}}$.

It follows immediately that

$X^2 = v + \emptyset\sqrt{2v}$ hence $X^2 = v(1 + \emptyset\sqrt{\frac{2}{v}})$

From standard normal theory $E\left(\frac{(X^2-v)}{\sqrt{2v}}\right) = 0$ and $V\left(\frac{(X^2-v)}{\sqrt{2v}}\right) = 1$

This implies that $E(\emptyset) = 0$ and $V(\emptyset) = 1$.

$\mathrm{E}(\emptyset^3) = \frac{8v}{(\sqrt{2v})^3} \Rightarrow \mathrm{E}(\emptyset^3) = \frac{\sqrt[3]{2}}{\sqrt{v}}$

$\mathrm{E}(\emptyset^4) = \frac{48v+12v^2}{4v^2} = \frac{12}{v} + 3$. In this case, as $v \to \infty$, $E(\emptyset) = 0$. This is evidence that any non-normal data with a chi-square distribution tends to normal as $v \to \infty$ under the influence of sufficient statistics. From the equation $X^2 = v\left(1 + \emptyset\sqrt{\frac{2}{v}}\right)$ we have

$\log(1+\mathrm{x}) = \mathrm{x} - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \ldots$

$\mathrm{Log}\, X^2 = log v + \log\left(1 + \emptyset\sqrt{\frac{2}{v}}\right) = \log v + \emptyset\sqrt{\frac{2}{v}} - \frac{\emptyset^2}{2}\frac{2}{v} + \frac{\emptyset^3}{3}\frac{\sqrt[3]{2}}{\sqrt[3]{v}} + \ldots$

$E(\mathrm{Log}\, X^2) = \log v - \frac{1}{v} + \frac{\sqrt[3]{2}}{\sqrt[3]{v}}\left(\frac{\sqrt[3]{2}}{\sqrt{v}}\right) \ldots$

$E(\mathrm{Log}\, X^2) = \log v + \omega\left(\frac{1}{v}\right)$.

Where ω is $\left[-1 + \frac{2^{2/3}}{v}\right]$

$E(\mathrm{Log}\, X^2 - \log v)^2 = E[\frac{\emptyset^2}{2}\frac{2}{v} - 2\emptyset^3\frac{1}{v}\sqrt{\frac{2}{v}} + \text{higher order terms in } \emptyset]$

$E(\mathrm{Log}\frac{\mathrm{X}^2}{\mathrm{v}})^2 = \frac{2}{v} - \frac{\sqrt[3]{2}}{\sqrt{v}}\frac{\sqrt[3]{2}}{\sqrt[3]{v}} + \text{powers of reciprocals of } v$

$V(\log X^2) = \frac{2}{v} + \omega\left(\frac{1}{v}\right)$

Let us come back to the estimation of $\hat{\lambda}$. $\hat{\lambda} = X^2$ has 2r degrees of freedom. This is to say that $v = 2r$. For this reason

$E(\mathrm{Log}\hat{\lambda}) = \log 2\mathrm{r} + \log\lambda - \log 2r + \omega\left(\frac{1}{r}\right)$

$E(\mathrm{Log}\hat{\lambda}) = \log\lambda + \omega\left(\frac{1}{r}\right)$

$V(\log\hat{\lambda}) = \frac{2}{2r} + \omega\left(\frac{1}{r}\right)$

$V(\log\hat{\lambda}) = \frac{1}{r} + \omega\left(\frac{1}{r}\right)$

$\sigma(\hat{\lambda}) = \sqrt{\frac{1}{r} + \omega\left(\frac{1}{r}\right)}$ (the general model has minimum variance)

Relating gamma distribution with parameter λ and r to chi-square, we have

$P[V \leq v] = P[X^2 \leq 2\lambda V]$

Since $X^2 = 2\lambda V = 2r\frac{\lambda}{\hat{\lambda}}$

$\hat{\lambda} = {}^{2\lambda r}/_{X^2}$

Noting that $2\hat{\lambda}V$ has a chi-square distribution with 2r degrees of freedom, an exact (1- α)100 confidence interval for $\hat{\lambda}$ is given by $P[X^2_{1-\alpha} < X^2 < X^2_{\alpha}] = 1 - 2\alpha$. Putting $2\lambda V$ for X^2 we have

$$P[X^2_{1-\alpha} < 2\lambda V < X^2_{\alpha}] = P\left[\frac{X^2_{1-\alpha}}{2V} < \lambda < \frac{X^2_{\alpha}}{2V}\right] = 1 - 2\alpha$$

For this reason, $1 - 2\alpha$ for λ has the end points at $\frac{X^2_{1-\alpha}}{2V}$ lower point and $\frac{X^2_{\alpha}}{2V}$ upper point.

At V large, $v \to \infty$, $X^2_{\alpha,v}$ tend to normal function hence

$$X^2_{\alpha,v} = v[1 - \frac{2}{9v} + Z_{\alpha}\sqrt{\frac{2}{9v}}]^3$$

It should be noted that non normal data follow normality assumptions.

Efficiency of the model

The general function for analyzing censored data is

$$f(\pi_i) = \lambda^r e^{-\lambda V}$$

where V is the total number of patients observed in the study. The aim is to estimate λ. From equation

$$\log f(\pi_1, \pi_2, \pi_3, \ . \ . \ . \ \pi_{r+1}) = r log\lambda - \lambda V + C$$

$\frac{dL}{d\lambda} = \frac{r}{\lambda} - V$. At $\frac{dL}{d\lambda} = 0$,

$$\hat{\lambda} = \frac{r}{V} \qquad (1)$$

Equation 1 is the optimal approach for estimating the hazard, this is because V is the total number of both observed values and censored values; $\hat{\lambda}$ is the likelihood of λ, where λ is the hazard. The mean time to death $\mu = \lambda^{-1}$. The maximum likelihood estimate $\hat{\mu}$ is expected to be the best linear unbiased estimator. Since $\mu = \lambda^{-1}$ is the mean time to death and maximum likelihood estimates are invariant under one-to-one transformations, we have

$$\hat{\mu} = \frac{V}{r}$$

For this reason, the minimum variance unbiased estimator for μ is important to obtain. Denote $\tilde{\mu}$ to be the minimum variance unbiased estimator of μ. For $\tilde{\mu}$ to be an unbiased estimator,

$$\tilde{\mu} = \sum_{i=1}^{r+1} a_i(t_i + S)$$

where $\sum_{i=1}^{n} a_i = 1$ is subject to constraint. $a_i = \frac{1}{n}$ for all i. For this reason

$$\tilde{\mu} = \frac{V}{r}$$

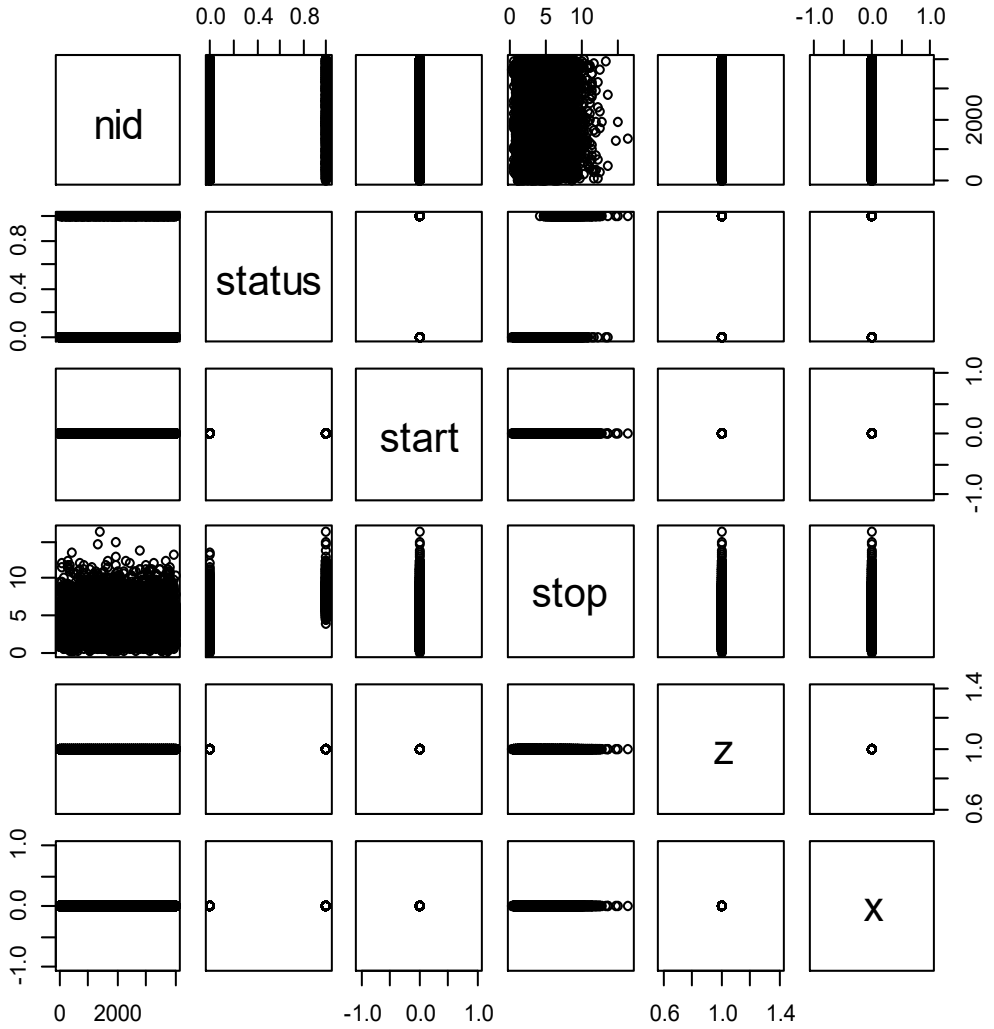
the maximum likelihood and minimum variance unbiased estimator of μ coincide as an exponential distribution function, since 1 is exponential. In estimation theory, $\frac{V}{r}$ is the best linear unbiased estimator (BLUE) for μ. One thing about BLUE is that, it is not generally preserved under inversing. For this reason, $\frac{r}{V}$ is not BLUE for λ.

Rao–Cramer lower bound efficiency is

$$Var(\tilde{\mu}) \geq \frac{\mu^2}{I(\emptyset)}$$

(3) is the efficiency function, where $Var(\tilde{\mu})$ is the minimum variance unbiased estimator, μ^2 is the square of the expectation and $I(\emptyset)$ is the Fisher information.

Figure 1 model simulation



A sample of 111 patients with breast cancer was obtained from the Korli-Bu Teaching Hospital in Accra Ghana in 2013. Patients times of entry into the hospital for medical checkup were recorded as well as their time of death (failure), time of absence from the hospital for treatment, time of relocating to a different community. The survival times of patients during data collection were also recorded. The data is made up of 807 censored observations consisting of 80% of the total observation and 204 failure observations representing 20% of the total observations were used for simulation. In Figure 1, *nid* represents a vector of unique identifier that identifies the individual patients who report for treatment. This means that, individual identity is maintained in the simulation process just as it is present in the original data. For this

reason, the model accounts for the censored values even though their stochastic realizations are masked. This means that simulation does not clump up patients' survival time but each patient's survival time is simulated according to the specified number of observations. The plot of the model captured patients survival status which represent a logical value indicating whether the survival time corresponds to the event (status = 1) or the survival time is censored (status = 0). From the plot, statuses are separated from each other. Censored status (status = 0) is shown in the upper corner and the event status is at the bottom of the status box. The status box is located at the right hand side of the status caption horizontally. Since clinical data are mostly multiple censored data the general proposed model $f(\pi_i) = \lambda^r e^{-\lambda V}$ is adequate for all types of censoring data. From the simulation plot, since all patients are allowed to start from a common point at time zero, the starting time is indicated on the plot. All patients are arranged to start at time zero to ensure normalized spacing t_i of patients' survival times. On the plot, it is obvious that patients' survival starting time in the simulation process is still maintained at zero. This is seen at the starting boxes horizontal to the start box. For this reason, the model accounted for reduction of bias likely to emanate from staggered entry that makes statistical analysis difficulty. The time the study is terminated is captured in the plot of the simulation. This corresponds to the censoring time. Censoring is the time the study is terminated. From the plot, the first box from the right on the horizontal stop box in indicated black at time zero but begins to fade to the terminating. From the simulation plot, symbol *Z* indicates whether there is an individual heterogeneity. Heterogeneity is the variation in individual survival times among the patients. The model does not give room for heterogeneity. This is indicated by the horizontal dark line.

DISCUSION

The time patients report for treatment are staggered and this makes survival entry time for patients' paucity. The resulted censored values make parameter estimation deviate from standard methods of survival distributions. In literature, methods like imputations, order statistics, Kaplan Meier and others were propounded to curb missing values or censored data. The magnitude of bias estimation in censored data

still persist. The robust model in this work minimize bias by putting all patients into a starting entry point of zero. For this reason, staggered entry points are removed making interval estimation easier in sub recursive intervals. From the model, censored data with extreme values and non-normal data are made to follow normality. Rao Cramer lower bound efficiency give the variance of the expected mean minimum variance unbiased estimator.

Many researchers have done much work on censored data estimations but staggered entry of patients distorts the robustness of their models.

One of the major contributions of this work is the proposed generalized model. The model was developed with respect to staggered entry of patients that poses estimation problems in censored data estimations. The model is useful when data ordering before estimation is not important. $f(\pi_i) = \lambda^r e^{-\lambda V}$ where V is the total number of patients observed in the study, λ is the scale parameter and r is the shape parameter. This is useful for estimating survival time for both independent, dependent continuous censored data. This general model is new in survival literature and is a contribution to knowledge in the subject. The generalized model proposed has specific properties such as efficiency, minimal standard error, robustness, consistency and asymptotic normality. Estimating hazard from the generalized model, with V which is the sum of both the censored time and uncensored survival time, enable full representation of the data and hence no loss of vital information. For this reason, the proposed model is efficient at representing maximum output and has a minimal standard error for estimating censored data. In the Cox Proportional model (Partial likelihood) the baseline hazard function is omitted in estimation, making parameters estimated only partially representative, inefficient and having larger standard errors. With large sample size, what is remarkable about the generalized proposed model is that it ensures gain is robustness, consistency and it is asymptotically normal. With this behavior of the distribution, the Cox model does not depend on the actual values of the event time but rather, it is fully based on the ranks of the survival times; any change in the coefficient β cannot alter the estimated results. Cox model depends on both the numerical values and the positions of the

survival times in predetermined intervals: any change in the coefficient β leads to a proportional change in the result.

References

1. Afifi, A. A., & Azen, S. P. (1972). *Statistical analysis: A computer oriented approach* (2nd ed.). Academic Press.

2. Barlow, R. E., & Proschan, F. (1965). *Mathematical theory of reliability*. John Wiley & Sons.

3. Boahen, E. (2014). Exponential model in clinical efficacy of *Cryptolepis sanguinolenta* on falciparum malaria treatment. *International Journal of Science and Technology*, *2*(7), 296–307.

4. Breslow, N. (1974). Covariance analysis of censored survival data. *Biometrics*, *30*(1), 89–99. https://doi.org/10.2307/2529620

5. Chakraborty, S. (2015). Generating discrete analogues of continuous probability distribution. *Journal of Statistical Distributions and Applications*, *2*(1), 1–11. https://doi.org/10.1186/s40488-015-0028-6

6. Clark, T. G., Bradburn, M. J., Love, S. B., & Altman, D. G. (2003). Survival analysis part I: Basic concepts and first analyses. *British Journal of Cancer*, *89*(2), 232–237. https://doi.org/10.1038/sj.bjc.6601118

7. Crowley, J., & Hu, M. (1977). Covariance analysis of heart transplant survival data. *Journal of the American Statistical Association*, *72*(357), 27–36. https://doi.org/10.1080/01621459.1977.10479903

8. Hong, J. S. (2009). Modified Weibull distribution in reliability. *Applied Sciences (APPS)*, *11*, 86–94. http://www.mathem.pub.ro/apps/v11/a11-6.pdf

9. Kasia, S., Nicholas, J., & White, S. (2006). Considerations in the design and interpretations of antimalarial drug trials in uncomplicated falciparum malaria. *Malaria Journal*, *5*(1), 1–12. https://doi.org/10.1186/1475-2875-5-6

10. Mudholkar, G. S., Srivastava, D. K., & Freimer, M. (1996). A generalization of the Weibull distribution with applications to the analysis of survival data. *Journal of the American Statistical Association*, *91*(436), 1575–1583. https://doi.org/10.1080/01621459.1996.10476725

11. Olshansky, S. J., & Carnes, B. A. (1997). Ever since Gompertz. *Demography*, *34*(1), 1–15. https://doi.org/10.2307/2061656

12. R Foundation for Statistical Computing. (n.d.). *R: A language and environment for statistical computing*. https://www.R-project.org/

13. Scholz, F. (2008). *Inference for the Weibull distribution*. University of Washington. https://www.stat.washington.edu/fritz/DATAFILES498B2008/WeibullBounds.pdf

14. Tsodikov, A. (2003). Semiparametric models: A self-consistency approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *65*(3), 759–774. https://doi.org/10.1111/1467-9868.00412