Scientific Research Publishing

Open Journal of Statistics, 2025, 15(3), 312-322
https://www.scirp.org/journal/ojs
ISSN Online: 2161-7198
ISSN Print: 2161-718X

Effects of Censoring Levels on Survival Data Analysis: Big Data

Evans Manjoro[1], Isheanesu Munyira[1], Charles Chimedza[2]

[1]Department of Mathematics and Computational Sciences, University of Zimbabwe, Harare, Zimbabwe
[2]School of Statistics and Actuarial Science, University of Witswaterand, Johannesburg, South Africa
Email: James.m.gregory@ttu.edu

How to cite this paper: Manjoro, E., Munyira, I. and Chimedza, C. (2025) Effects of Censoring Levels on Survival Data Analysis: Big Data. *Open Journal of Statistics*, **15**, 312-322.
https://doi.org/10.4236/ojs.2025.153016

Received: May 29, 2025
Accepted: June 27, 2025
Published: June 30, 2025

Abstract

In addition to non-normality, censoring is one of the characteristics of survival data. All traditional procedures and models take into consideration this censoring characteristic in relation to survival data analysis. However, no studies have been done on the effect of censoring levels in survival data analysis. The main objective of this paper is to look at the effect of censoring levels in survival data analysis in relation to big data. Data of sizes n = 10,000, n = 50,000 and n = 100,000 were simulated each at censoring levels of p = 0.1, p = 0.5 and p = 0.9. For comparison sake, also small/moderate sized survival datasets were also simulated. Censoring levels had a low effect on small/moderate sized datasets and had a significant effect on big datasets. This was depicted by the plots of survivor function. Visually, it was evident that as the level of censoring increases, there is a tendency to overestimate survival prospects. Model fit was much better for small/moderate datasets as compared to model fit for big datasets. This supports the idea of many researchers that traditional survival statistical models are inferior when handling big data. Surprising, the model fit for high censoring level (p = 0.9) had a much better fit both on small/moderate and big datasets.

Keywords

Survival, Censoring, Big Data

1. Introduction

Historically, the event of primary interest in survival analysis is patient morbidity, but active research in the field over the past few decades has provided applications spread across many domains including biostatistics, sociology, economics, demography, and engineering [1]-[4]. In addition to survival times, additional information on covariates has been used to evaluate their importance in predicting the survival

probability of a given observed individual. This now entails modeling of survival data. One of the main challenges for such time-to-event data is that usually there exist censored instances, i.e., the event of interest is not observed for these instances due to either the time limitation of the study period or losing track during the longitudinal observation period [5]. This issue of censoring is a significant challenge of survival analysis which all methodologies have to take into consideration.

In survival analysis, censoring typically takes the form of random right, left and interval censorship. Survival analysis, an important sub-field of statistics, provides various mechanisms to handle such censored data problems that arise in modeling such data. Traditionally, mostly statistical models were developed to overcome this censoring issue, however no studies have been done to determine the effect of censoring level. Also studies do not even mention the extend of censoring occurrences.

In 1972, Sir David Cox read the paper "Regression models and life tables" to the Royal Statistical Society [6]. In this seminal paper, Cox (1972) presented the proportional hazards model, which specifies that the conditional hazard function of failure time given a set of covariates is the product of an unknown baseline hazard function and an exponential regression function of covariates [6]. This proportional hazards model of Cox (1972) has been the most popular survival analysis statistical model and its main concept is to analyze the relationships between multiple covariates and survival time. For classical/traditional problems with many more observations than predictors, the Cox model performs well [7]. The main objective of this paper is to determine the performance of this Cox model under various levels of censoring for both small/moderate and big datasets.

2. Literature

The term "big data" was used for the first time in a 1997 article by NASA researchers Michael Cox and David Ellsworth. They claimed that the rise of data was becoming an issue for current computer systems. The term "Big Data" has several definitions. The term "Big Data" refers to the evolution and use of technologies that provide the right user at the right time with the right information from a mass of data that has been growing exponentially for a long time in our society [8]. Five characteristics have been used to define big data, also known as 5V's (Volume, Variety, Velocity, Veracity and Value). Volume refers to the size of the data sets that need to be analyzed and processed, which are now larger than the ones for traditional data sets. Velocity refers to the speed at which the data is accumulated. Social media messages go viral in seconds. Variety refers to the multiplication of the types of data managed by an information system. Veracity refers to the assurance of quality/integrity/credibility/accuracy of the data. Value refers to the important feature of the data which is defined by the added-value that the collected data can bring to the intended process, activity or predictive analysis/hypothesis [9]. Of these five characteristics our main focus is on volume (many instances).

Survival data have two important special characteristics: (1) survival times are non-negative and consequently are usually positively skewed and (2) typically some

subjects (units of observation) have censored survival times that is survival times of these subjects are not observed. However, of these two special characteristics of survival times, the one which renders most statistical procedures redundant is the issue of censoring. There are three types of censoring: right censoring, left censoring and interval censoring. Right censoring occurs when a subject leaves the study before an event occurs, or the study ends before the event has occurred and it is the most prevalent type of censoring. Left censoring is when it is known that the failure occurs some time before the recorded follow-up period. That is, the actual survival time is less than the observed survival time. With interval censoring it is known that the event occurs between two times, but the exact time of failure is not known.

Survival data are normally represented by a triple of variables (X, T, δ), where X is the feature vector, and δ is an indicator variable. $\delta = 1$ if T is the time to the event of interest and $\delta = 0$ if T is the censored time; for convenience, T is usually name the observed time [4]. In summarizing survival data, there are three functions of central interest, namely the survivor function, the hazard function, and the cumulative hazard function [10].

Suppose T is a non-negative random variable representing the time until some event of interest. Now suppose that this random variable, T has a probability distribution with underlying probability density function $f(t)$. It is usually assumed that T is continuous and its distribution function is given by

$$F(t) = P[T < t] = \int_0^t f(u)\,\mathrm{d}u \tag{1}$$

This function is also called the cumulative incidence function, since it summarizes the cumulative probability of death occurring before time t [10]. The survival function, which is denoted by $S(t)$ is the probability that a subject survives beyond time t and it is given by

$$S(t) = P(T > t) = 1 - F(t) \tag{2}$$

In relation to application in medical domain, the survival function is always a non-increasing function.

One other function commonly used in survival analysis is the hazard function ($h(t)$), which is also known as the force of mortality, the conditional failure rate, or the instantaneous death rate [11]. Hazard function is defined as the probability of dying at time t conditional on subject having survived to time t and Mathematically defined as

$$\begin{aligned} h(t) &= \lim_{\delta t \to \infty} \frac{P(t \le T < t + \delta t \mid T \ge t)}{\delta t} \\ &= \lim_{\delta t \to \infty} \frac{F(t + \delta t) - F(t)}{\delta t \cdot S(t)} \\ &= \frac{f(t)}{S(t)} \end{aligned} \tag{3}$$

In contrast to survival function which is always non-increasing, the hazard function can take any shape. The cumulative hazard function is given by:

$$\begin{aligned} H(t) &= \int_0^t h(u)\,\mathrm{d}u \ (t>0) \\ &= -\ln\left(1-F(t)\right) \\ &= -\ln\left(S(t)\right) \end{aligned} \tag{4}$$

Cox Proportional Hazards Model

The general Cox proportional hazard model is given by

$$h(t,\boldsymbol{x}) = \exp\left(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}\right)h_0(t) \tag{5}$$

where $h_0(t)$ is the baseline hazard function which depends on time but not covariates and $\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x} = \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi}$. The model is semi-parametric because the baseline can take any form and the covariates enter model linearly. Fitting the proportional hazards model given in equation 5 to an observed set of survival data entails estimating the unknown coefficients of the explanatory variables $X_1, X_2, \cdots, X_p$ in the linear component of the model, $\beta_1, \beta_2, \cdots, \beta_p$ The β-parameter estimates can then be found by maximizing this log-likelihood function using numerical methods and this maximization is usually accomplished using the Newton-Raphson procedure.

3. Methods

In real life, most of the survival data sets are small/moderately sized and traditional statistical models work well. However, for this paper, consideration is placed on big datasets in relation to size (many instances-volume). The study will use simulated data.

When conducting simulation studies to evaluate the performance of new and existing statistical methods for analyzing survival data, one is required to simulate event times under a known data generating model [1]. Traditionally, a common approach has been to make simplifying parametric assumptions about the distribution of the event times [1]. In survival analysis the most prominent model is the Cox proportional hazard model and it doesn't assume any distribution of the event times. Thus we have a problem with the above suggested approach because it contradicts a key advantage of the Cox model that is the ability to leave the distribution of the baseline hazard unspecified. However, no standard method exists for simulating durations directly from its data generating process because it does not assume a distributional form for the baseline hazard function [12]. [12] proposed a method that generates a baseline hazard function at random by fitting a cubic spline to randomly drawn points. This produces a wide variety of shapes for the baseline hazard, including those that are unimodal, multimodal, monotonically increasing or decreasing, and other shapes. Survival times drawn from this function match the Cox model's inherent flexibility and improve the simulation's generalization. However, one of the main disadvantages of this approach is on the covariates. It only simulates duration times with continuous covariates, however in many situations covariates are factors.

Three factors of different sizes ($n = 50$, $n = 80$, $n = 100$, $n = 10{,}000$, $n = 50{,}000$ and $n = 100{,}000$) were simulated. The first factor has two (2) factor levels, the second factor has three (3) factor levels and the last one has four (4) factor levels. The reason for using factors is that most survival studies use continuous covariates and we wanted to deviate from this norm. We also considered censoring levels of 10%, 50% and 90% to capture the effects of low, moderate and high censoring. [13] considered three levels of censoring, 10%, 25%, and 70%, representing relatively light, moderate, and relatively heavy censoring, respectively on random forests for survival data. Using the *coxed function* (Duration-Based Quantities of Interest for the Cox Proportional Hazards Model) we then simulated the survival data (see Appendix).

After the simulations, the data was ready for analysis and for commands see Appendix. Descriptive measures, the histogram and survivor functions were used. The Proportional Hazard Model (PHM) was fitted to all the datasets and evaluation metric Concordance was used to determine the goodness of fit of the fitted model.

4. Results

Figure 1 shows histograms for all censoring levels ($n = 10{,}000$) that is at $p = 0.1$, $p = 0.5$ and $p = 0.9$. The lower histogram is for the actual survival times and the upper one is for the censored survival times. The notable thing for these histograms is that for $p = 0.5$ the shapes for both survival and censored observations are the same (Figure 2, Figure 3). This similarity shape, $p = 0.5$ was also noted for other big datasets that is $n = 50{,}000$ and $n = 100{,}000$ as shown in appendices Figure C1 and Figure C2. As shown in Figure C3 and Figure C4 for moderate sized dataset ($n = 50$, $n = 100$), there is no similarity in shape as compared with big datasets. This difference in shape at $p = 0.5$ was also noted for other moderate datasets, $n = 80$ and $n = 100$.

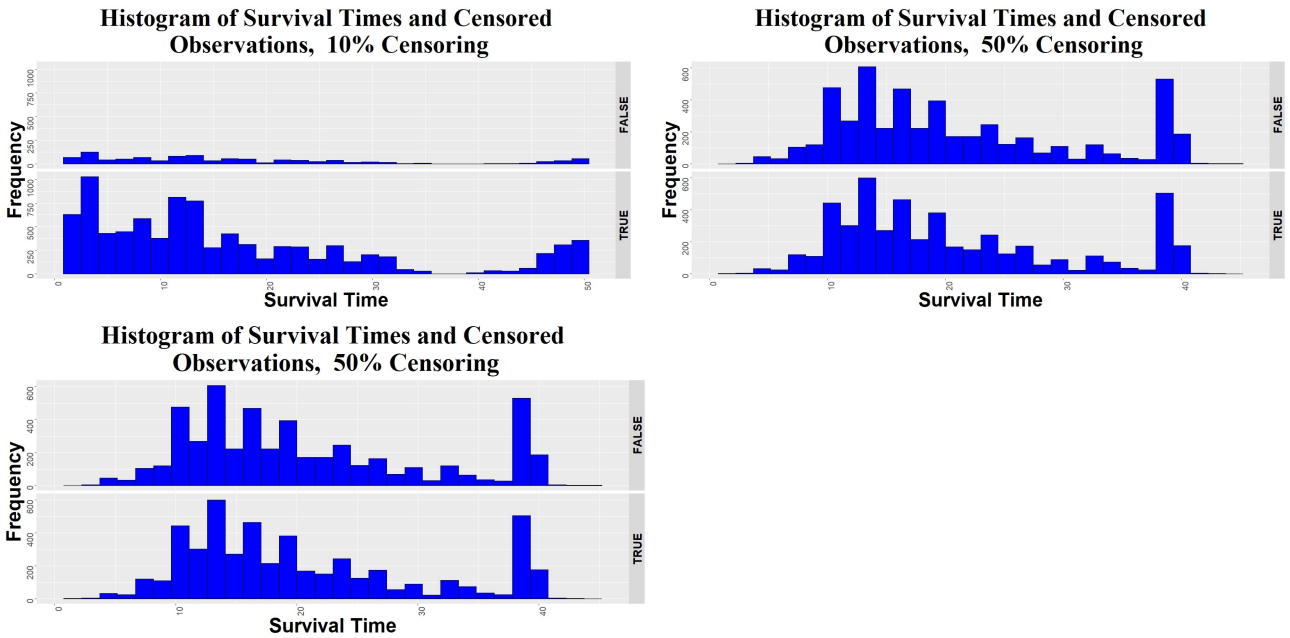


Figure 1. Histograms of survival times ($n = 10{,}000$) for 10%, 50% and 90% censoring levels, big datasets.

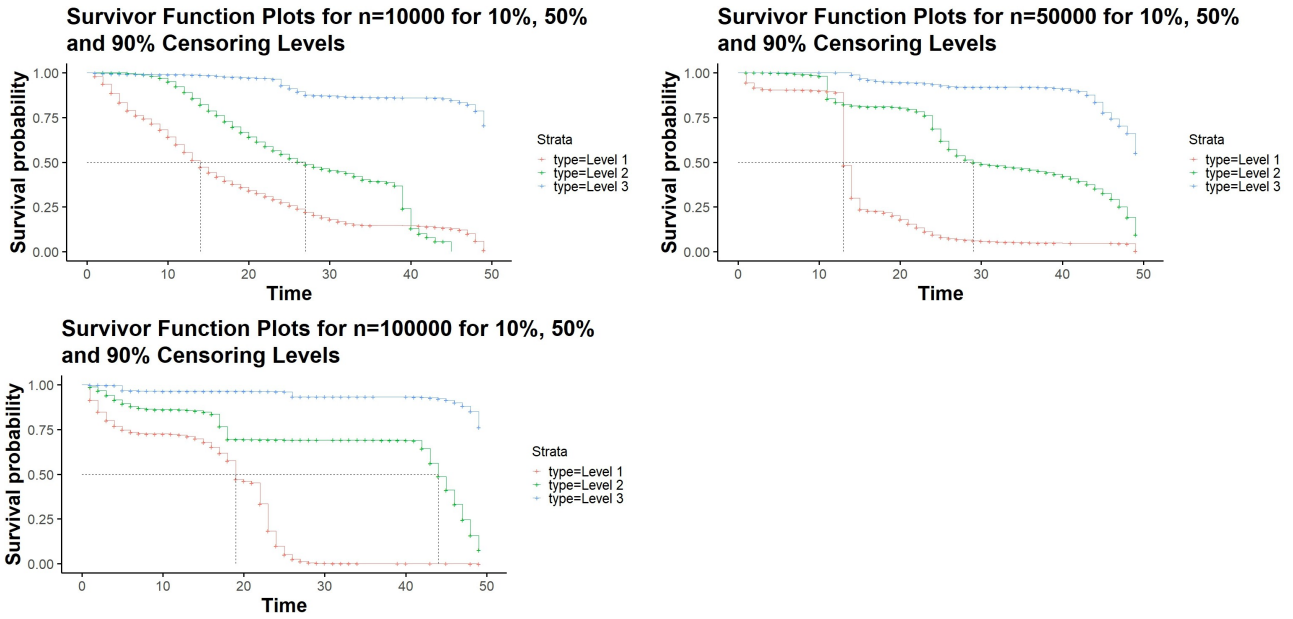


Figure 2. Survivor functions, big datasets (n = 10,000, n = 50,000 and n = 100,000).

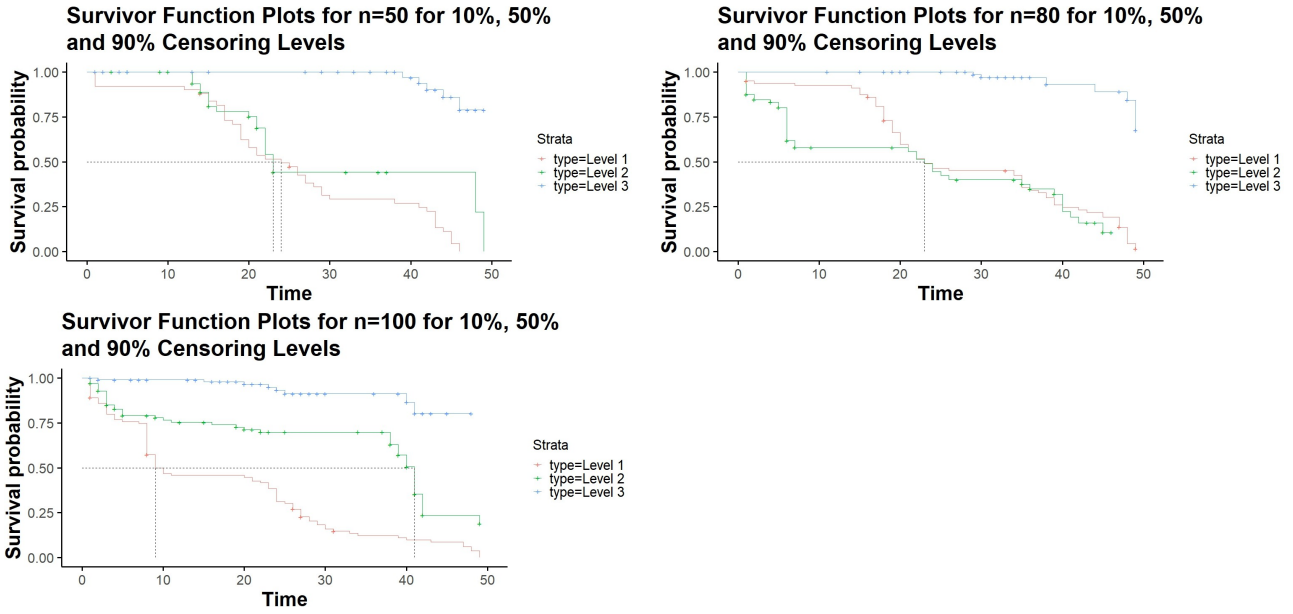


Figure 3. Survivor functions, small/moderate datasets (n = 50, n = 80 and n = 100).

Figure 2 and Figure 3are survivor function plots at different levels of censoring for big datasets (n = 10,000, n = 50,000, n = 100,000) and moderate sized datasets (n = 50, n = 80, n = 100) respectively. For big datasets there are marked differences in all survivor functions and for the moderate sized datasets there are no differences at lower levels of censoring that's at p = 0.0.1 and p = 0.05. However, for both small/moderate and big datasets, there is a marked difference at p = 0.9 compared to other censoring levels. From both graphs it is quite evident that the level of censoring affects the estimation of survivor function especially at high levels of censoring. At a high level of censoring, there is an overestimation of survivor func-

tion to the extent that it is impossible to estimate measures such as median survivor function and upper percentiles of survivorship.

Table 1 shows model fit Concordance statistics. Moderate datasets have a better fit as compared to big datasets which had a poor fit of around 50%. For both datasets there is not much difference in fit at lower levels of censoring, that is, at $p = 0.1$ and $p = 0.5$. However, there is a marked change in model fit at a high level of censoring, $p = 0.9$. The model fit has improved quite a lot at this censoring level, and this is a surprising result.

Table 1. Model fit concordance statistics.

	Concordance Statistics		
n	$p = 0.1$	$p = 0.5$	$p = 0.9$
50	0.629	0.762	0.838
80	0.614	0.687	0.763
100	0.655	0.653	0.857
10,000	0.542	0.542	0.568
50,000	0.525	0.511	0.549
100,000	0.583	0.555	0.583

5. Conclusions

The magnitude of censoring was something researchers didn't consider when they came up with different methodologies for handling survival data. However, from this study, it is evident that the censoring level affects the results of survival analysis, especially when working with big datasets. There is an overestimation of the survivor function when censoring levels are quite high to the extent that measures such as median survival time and higher percentiles survival times can't be estimated. Also the Cox Proportional hazard model performance is quite poor as shown by low Concordance statistics. It then calls for other methodologies for handling big survival datasets. To tackle practical concerns which can't be addressed by statistical models, some related works have adapted machine learning methods to solve survival analysis problems and in this field researchers have developed more sophisticated and effective algorithms which either complement or compete with the traditional statistical methods [5].

Censoring is something we can't do without when dealing with survival data. However, it is also something researchers can control when they set their studies correctly. Research practitioners should exhaustively consider all tracking variables on participants so as to avoid censoring due to loss of follow-up. Also they should come up with realistic study time so that by the end of the study, the event of interest would have probably occurred. For big data, we recommend the use of machine learning approaches such as survival trees, random forests, and neural networks for survival data.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Brilleman, S.L., Wolfe, R., Moreno-Betancur, M. and Crowther, M.J. (2021) Simulating Survival Data Using the Simsurv R Package. *Journal of Statistical Software*, **97**, 1-27. https://doi.org/10.18637/jss.v097.i03 https://www.jstatsoft.org/index.php/jss/article/view/v097i03

[2] Collett, D. (2003) Modelling Survival Data in Medical Research. 2nd Edition, Chapman & Hall/CRC Texts in Statistical Science, Taylor & Francis. https://books.google.co.zw/books?id=4t3-GWDKDRQC

[3] Heckman, J.J. and Robb, R. (1985) Alternative Methods for Evaluating the Impact of Interventions. In: Heckman, J.J. and Singer, B.S., Eds., *Longitudinal Analysis of Labor Market Data*, Cambridge University Press, 156-246. https://doi.org/10.1017/ccol0521304539.004

[4] Lee, E.T. and Wang, J.W. (2003) Statistical Methods for Survival Data Analysis. Wiley. https://doi.org/10.1002/0471458546

[5] Wang, P., Li, Y. and Reddy, C. (2017) Machine Learning for Survival Analysis: A Survey. arxiv abs/1708.04649.

[6] Lin, D.Y. (2007) On the Breslow Estimator. *Lifetime Data Analysis*, **13**, 471-480. https://doi.org/10.1007/s10985-007-9048-y

[7] Simon, N., Friedman, J., Hastie, T. and Tibshirani, R. (2011) Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent. *Journal of Statistical Software*, **39**, 1-13. https://doi.org/10.18637/jss.v039.i05

[8] Riahi, Y. and Riahi, S. (2018) Big Data and Big Data Analytics: Concepts, Types and Technologies. *International Journal of Research and Engineering*, **5**, 524-528. https://doi.org/10.21276/ijre.2018.5.9.5

[9] Hiba, J., Hadi, H., Hameed Shnain, A., Hadishaheed, S. and Haji, A. (2015) Big Data and Five V's Characteristics. 2393-2835. https://www.iraj.in/journal/journal_file/journal_pdf/12-105-142063747116-23.pdf

[10] Collet, D. (2015) Modelling Survival Data in Medical Research. Chapman & Hall/CRC Texts in Statistical Science, CRC Press. https://books.google.co.zw/books?id=Okf7CAAAQBAJ

[11] Dunn, O.J. and Clark, V.A. (2009) Basic Statistics. Wiley. https://doi.org/10.1002/9780470496862

[12] Harden, J.J. and Kropko, J. (2018) Simulating Duration Data for the Cox Model. *Political Science Research and Methods*, **7**, 921-928. https://doi.org/10.1017/psrm.2018.19

[13] Berkowitz, M., Altman, R.M. and Loughin, T.M. (2024) Random Forests for Survival Data: Which Methods Work Best and under What Conditions? *The International Journal of Biostatistics*, **20**, 315-345. https://doi.org/10.1515/ijb-2023-0056

Appendix A. SIMULATIONS, R (n = 10,000)

```
#Simulation of Factors #Two factor levels
>F2=sample.int(2,10000,replace=TRUE) #Three factor levels
>F3=sample.int(3,10000,replace=TRUE) #Four factor levels
>F4=sample.int(4,10000,replace=TRUE) #Covariates Matrix
>Fdata=cbind(F2,F3,F4)    #SURVIVAL DATA SIMULATIONS
>library(coxed)
>SimData_0.1=sim.survdata(T=49,X=Fdata,censor=0.1,num.data.frames=1)
>SimData_0.5=sim.survdata(T=49,X=Fdata,censor=0.5,num.data.frames=1)
>SimData_0.9=sim.survdata(T=49,X=Fdata,censor=0.9,num.data.frames=1)
#Correct Data Frame
>SimData_0.1_1=head(SimData_0.1$data,10000)
>SimData_0.5_1=head(SimData_0.5$data,10000)
>SimData_0.9_1=head(SimData_0.9$data,10000)
```

Appendix B. ANALYSIS COMMANDS, R

```
#Histogram
>library(ggplot2)
>ggplot(SimData_0.1_1,    aes(x=y,    color=failed))
+ggtitle("Histogram of Survival Times and Censored Observations, p=0.1") +
geom_histogram(fill="white", position="dodge")+ theme(legend.position="top")
>ggplot(SimData_0.5_1,    aes(x=y,    color=failed))
+ggtitle("Histogram of Survival Times and Censored Observations, p=0.5") +
geom_histogram(fill="white", position="dodge")+ theme(legend.position="top")
>ggplot(SimData_0.9_1,    aes(x=y,    color=failed))
+ggtitle("Histogram of Survival Times and Censored Observations, p=0.9") +
geom_histogram(fill="white", position="dodge")+ theme(legend.position="top")
#Survivor Function
>library(survminer)
Sim_All <- rbind((cbind(SimData_0.1_1, type = 'Level 1')), (cbind(Sim-
Data_0.5_1, type = 'Level 2')), (cbind(SimData_0.9_1, type = 'Level 3')))
>KM_ALL= survfit(Surv(y, failed) ~ type, data=Sim_All) #Cox Proportional
Hazard Model Fit
>cox _0.1= coxph(Surv(y, failed) ~ factor(F1) + factor(F2)
+ factor(F3), data = SimData_0.1_1)
>cox _0.5= coxph(Surv(y, failed) ~ factor(F1) + factor(F2)
+ factor(F3), data = SimData_0.5_1)
>cox _0.9= coxph(Surv(y, failed) ~ factor(F1) + factor(F2)
+ factor(F3), data = SimData_0.9_1)
>summmary(cox_0.1)
>summmary(cox_0.5)
>summmary(cox_0.9)
```

Appendix C. OTHER OUTPUTS

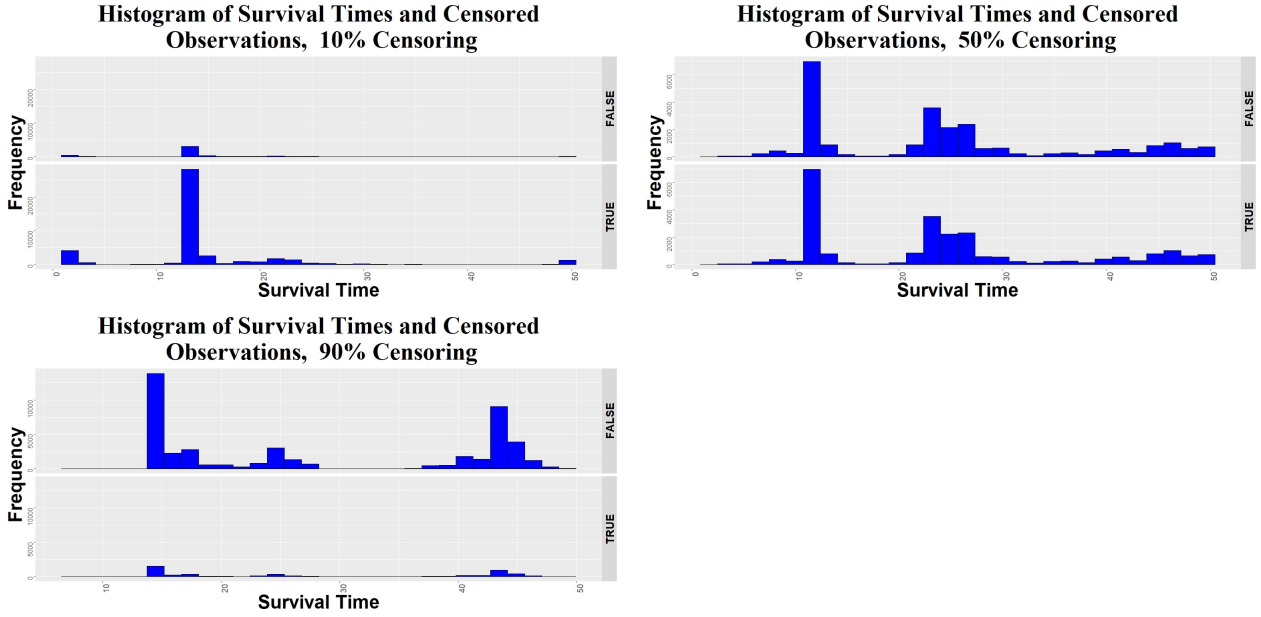


Figure C1. Histograms of survival times (n = 50,000) for different censoring levels, big datasets.

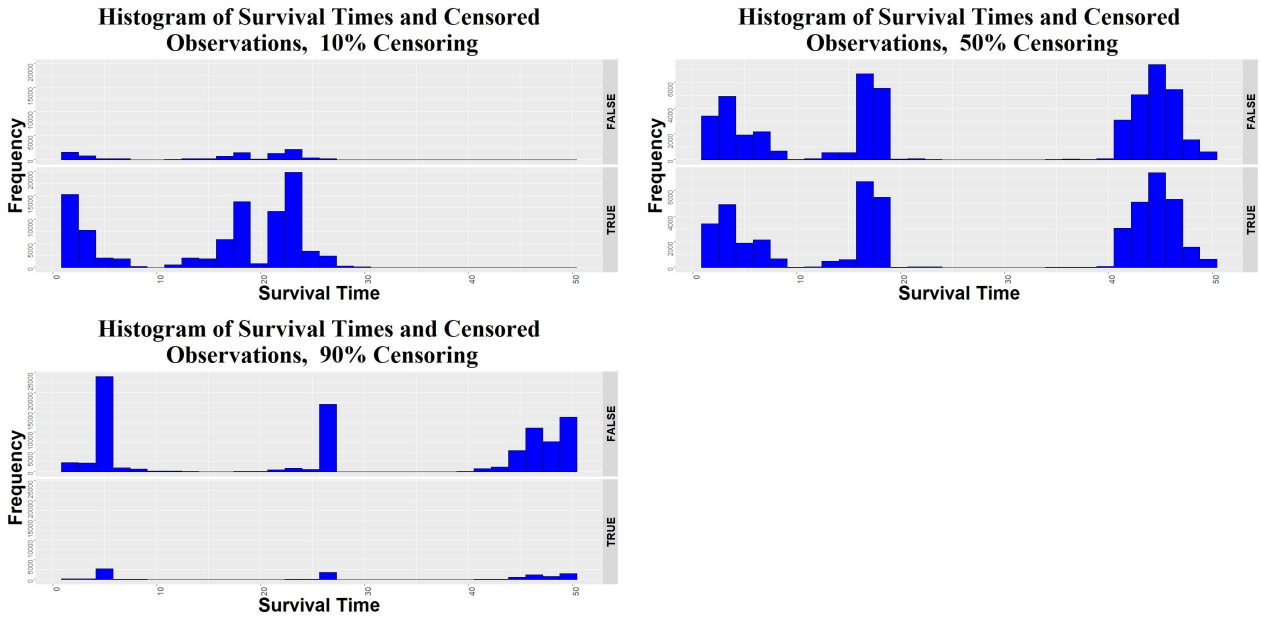
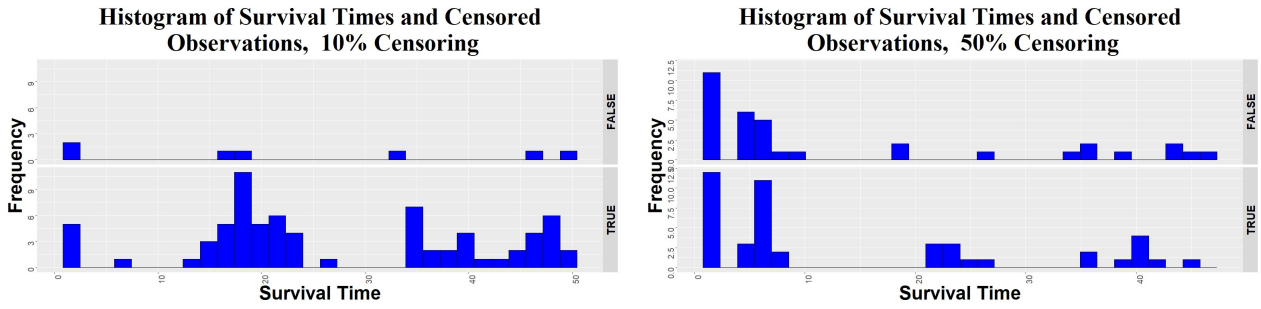


Figure C2. Histograms of survival times (n = 100,000) for different censoring levels, big datasets.



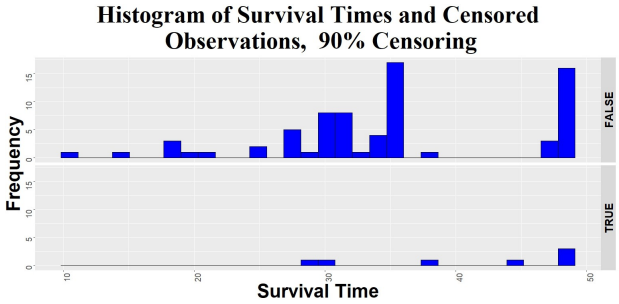


Figure C3. Histograms of survival times ($n = 80$) for different censoring levels, moderate datasets.

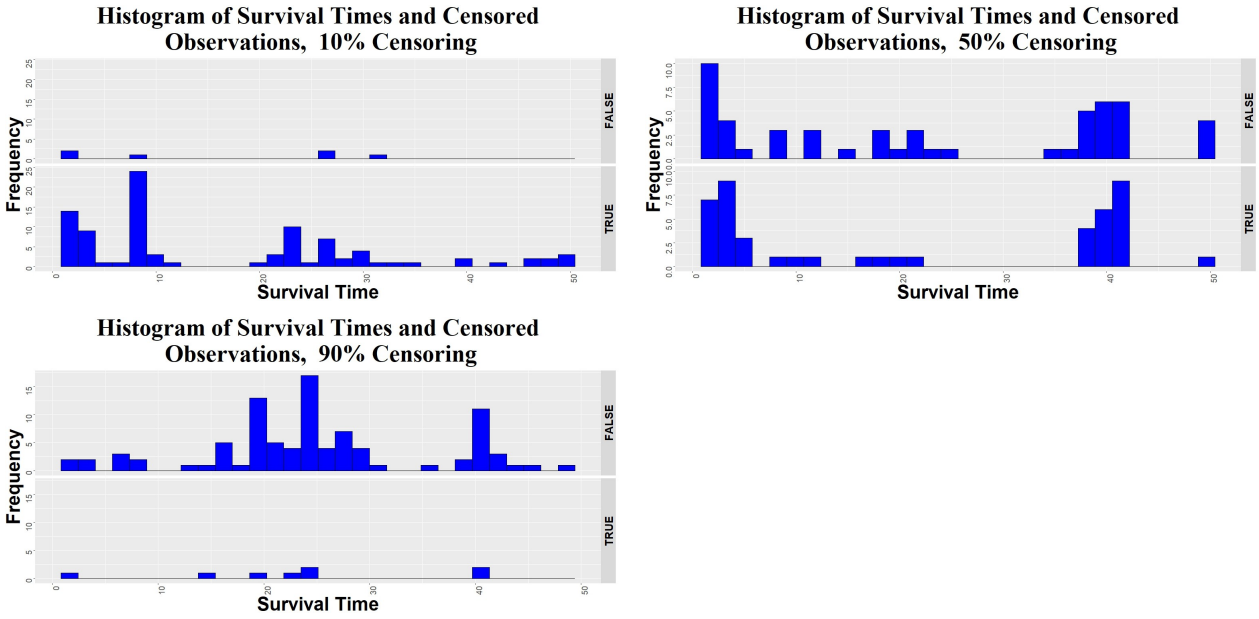


Figure C4. Histograms of survival times ($n = 100$) for different censoring levels, moderate datasets.