STATISTICA, anno LXXVI, n. 4, 2016

SEMIPARAMETRIC MODELS WITH COVARIATES FOR LIFETIME DATA UNDER A GENERAL CENSORING SCHEME WITH AN APPLICATION TO CONTINGENT VALUATION

Nathan Bennett
Department of Statistics and Applied Probability, University of California Santa Barbara, CA, USA

Srikanth K. Iyer [1]
Department of Mathematics, Indian Institute of Science, Bangalore, 560012 India

Sreenivasa Rao Jammalamadaka
Department of Statistics and Applied Probability, University of California Santa Barbara, CA, USA

1. Introduction

Our aim in this paper is to estimate the lifetime distribution or its complement, the survival function, for data that is subject to middle censoring. Middle censoring occurs when a data point becomes unobservable if it falls inside a random interval. This is a generalization of left and right censored data and is quite distinct from the case of doubly censored data.

Middle censoring was first introduced by Jammalamadaka and Mangalam (2003) for non-parametric estimation of lifetime distributions. Middle censored data was analyzed in Iyer, Jammalamadaka and Kundu (2008) when the lifetimes are exponentially distributed, whereas Jammalamadaka and Mangalam (2009) study such censoring in the context of circular data. An example of middle censoring could be when a patient temporarily withdraws from a clinical study, but is later re-entered into the study. This could also happen if a patient is in a trial under continual evaluation and the monitoring equipment fails or power goes out for a period of time before being able to resume measurements.

In many situations there is auxiliary information about the subjects under study, in the form of covariates $\boldsymbol{Z}$, which may affect the lifetimes of individuals. This is natural since every machine is operated under different conditions, and different people will have different medical histories which affect their lifetimes. For example, a doctor might want to know how long a person with diabetes will live. Each person has a different age, weight, blood pressure, sex, family history, and many other factors affecting their overall health. Obviously, one must account for these covariates in a model in order for it to be effective.

[1] Corresponding Author. E-mail: srikiyer@gmail.com

The case of middle censoring with covariates has recently been examined. Jammalamadaka, Prasad, and Sankaran (2016) provide a general, semi- parametric approach to regression in the context of middle censoring with minimal constraint on the baseline lifetime distribution. Bennett, Iyer, and Jammalamadaka (2017) examines a fully parametric approach to middle censoring with covariates, where the baseline lifetime distribution is either gamma or Weibull distributed.

In this paper we consider two models that are commonly applied, namely the Cox proportional hazard model (PH) and the accelerated failure time model (AFT). The approach given here differs from Jammalamadaka, Prasad, and Sankaran (2016) in that more attention is paid to the baseline lifetime distribution. An alternative approach to estimating the baseline lifetime distribution is given and an additional covariate model is studied, namely the AFT model. This paper also differs from BIJ (2017) since it takes a much more general approach to estimating the baseline survival function and considers the semi-parametric Cox PH model.

The proportional hazards model was introduced by Cox (1972). Under this model, there is a common hazard function but each patient has a modifier which multiplies the hazard by a function of the covariates. Specifically, it is defined by

$$h(t|\mathbf{Z}) = h_0(t)\,\psi(\mathbf{Z}). \tag{1}$$

The baseline hazard function, $h_0(t)$, is usually considered to be a nuisance parameter and is not estimated. A common choice for the function of the covariates is $\psi(\mathbf{Z}) = \exp\left(\theta^T \mathbf{Z}\right)$. In this case, the baseline hazard function simplifies to

$$h_0(t) = h(t|\mathbf{Z} = \mathbf{0}). \tag{2}$$

Thus the PH model is a semi-parametric model. With the above choice of $\psi(\mathbf{Z})$, we can rewrite the model in terms of the survival function

$$S(t|\mathbf{Z}) = \left(S_0(t)\right)^{\psi(t)}, \tag{3}$$

where $S_0(t)$ is the survival function associated with the baseline hazard function $h_0(t)$. Looking at the natural logarithm of the ratio of hazards, one will notice a proportionality. Hence, this model is commonly referred to as the "proportional hazards" model. This is given by

$$\ln\left(\frac{h\left(t|\mathbf{Z}\right)}{h\left(t|\mathbf{Z}^*\right)}\right) = \theta^T\left(\mathbf{Z} - \mathbf{Z}^*\right). \tag{4}$$

One can see that for two individuals with different sets of covariates, $\mathbf{Z}$ and $\mathbf{Z}^*$, the proportional hazard of one individual relative to the other is constant over time.

To the best of the authors' knowledge, existing literature on this model does not estimate the baseline hazard function, $h_0(t)$, but only the parameter θ that reflects the impact of the covariates,$\mathbf{Z}$,on the lifetimes. Cox (1972) suggested using the following partial likelihood to estimate the regression coefficients

$$L(\beta) \propto \prod_{i=1}^{n}\left(\frac{\exp\left(\theta^T Z_i\right)}{\sum_{j\in R(T_i)}\exp\left(\theta^T Z_j\right)}\right). \tag{5}$$

While this is not the complete likelihood, it is shown to provide good estimates with a small loss of efficiency (Efron (1977)). This model has been further developed and refined to make it more accurate. We show how the baseline survival function can be estimated when the covariate distribution has mass at 0. This assumption is required to obtain an initial estimate for the baseline survival function. Otherwise there is an identifiability issue with the model. In the case of continuous covariates, one could use the data corresponding to covariate values around zero to obtain an initial estimate. The PH model is most commonly used in biology and medicine due to its flexibility and the baseline hazard function does not need to be known.

The Cox Proportional Hazard model has been studied in the case of right censoring (see for instance Cox (1972) and Efron (1977)). This model is heavily relied upon in the study of medical data and pharmaceutical studies. As such, a more general class of censoring mechanism, such as the middle censoring, would be beneficial in many of these applications.

Another popular model for survival analysis is the Accelerated Failure Time (AFT) model. Mathematically, this model is given by

$$S(t|\mathbf{Z}) = S_0\left(\exp\left(\beta^{\mathbf{T}}\mathbf{Z}\right)\, t\right), \tag{6}$$

where $\exp\left(\beta^{\mathbf{T}}\mathbf{Z}\right)$ is called the accelerating factor. The reason why this is called an *accelerated* failure time model is because the covariates, $\mathbf{Z}$, can either speed up or slow down the failure time of an individual.

A key contribution of this paper is a method which estimates the baseline survival function under a general censoring scheme as well as the effect of the covariates. In Sections 2 and 3 we derive the estimation procedure for the Cox Proportional Hazard model and the Accelerated Failure Time model respectively. In Section 4 we apply this procedure to the problem of contingent valuation.

2. Proportional Hazards Model

In this section, we discuss the problem of estimation of the baseline as well as the parameter of the lifetime distribution in the presence of middle censoring. Recall that the Cox PH model is given by

$$S\left(t|Z\right) = S_0\left(t\right)^{\exp(\theta Z)}, \tag{7}$$

where $S\left(t\right)$ is the survival function for a non-negative random variable. With this semi-parametric set-up, the density of lifetimes is given by

$$f\left(t|Z\right) = -\frac{\partial}{\partial t}S\left(t|Z\right) = f_0\left(t\right)\exp\left(\theta Z\right)\left[S_0\left(t\right)\right]^{\exp(\theta Z)-1}. \tag{8}$$

To the best of the authors' knowledge, previous research done with this model was only concerned with estimating the regression parameters, θ. The baseline survival function is treated as a nuisance parameter and is not estimated. It would be extremely beneficial to have a method that estimates both the regression parameters, θ, and the baseline survival function.

Let us denote the "actual" lifetimes of the n individuals by $t_1, \cdots, t_n$, and not all of them are observable. For each individual there is a random period of time $[\ell_i, r_i]$ for which the lifetime of the i^{th} individual is unobservable. Thus,

the actual lifetime is observed if $t_i \notin [\ell_i, r_i]$ and if $t_i \in [\ell_i, r_i]$ then only the interval is observed. The lifetimes are assumed to be independent and identically distributed (*i.i.d.*) from an unknown probability density function of the form given in (8). Additionally, the censoring intervals, $[L_1, R_1], \cdots, [L_n, R_n]$, are assumed to be *i.i.d.* from an unknown bivariate distribution function $G(\cdot, \cdot)$. Finally, the lifetimes and the censoring intervals are taken to be independent of each other, as is common in survival analysis.

For convenience of notations and without any loss of generality, let the following be the observed failure times

$$uncen = (t_1, \cdots, t_{n_1}) \tag{9}$$

and the following denote the observed middle censored data from (8) under the general censoring scheme described above

$$cen = ((l_{n_1+1}, r_{n_1+1}), \cdots, (l_{n_1+n_2}, r_{n_1+n_2}.)) \tag{10}$$

Then the full likelihood is given by

$$L(\theta) = \prod_{uncen} f(t_i|z_i) \prod_{cen} [S(l_i|z_i) - S(r_i|z_i)]. \tag{11}$$

Hence, the corresponding log-likelihood is

$$l_{full}(\theta) = l_{uncen}(\theta) + l_{cen}(\theta), \tag{12}$$

where

$$l_{uncen}(\theta) = \sum_{uncen} \ln(f_0(t_i)) + \theta \sum_{uncen} z_i + \sum_{uncen} (\exp(\theta z_i) - 1) \ln(S_0(t_i)) \tag{13}$$

and

$$l_{cen}(\theta) = \sum_{cen} \ln(S(l_i|z_i) - S(r_i|z_i)). \tag{14}$$

This requires estimation of the baseline survival function, $S_0(t)$, or equivalently the baseline density, $f_0(t)$, in order to estimate the covariate effect, θ. Estimating the survival function non-parametrically can be done by using the self-consistent estimator (SCE) or the non-parametric MLE (NPMLE) given in Jammalamadaka and Mangalam (2003) provided we could modify the data appropriately. One can estimate the baseline density function by fitting a smoothing spline to the estimate of the baseline survival function and differentiating it, but this can cause large errors in estimating the derivative of the log-likelihood. We can avoid this by writing out the derivative of the log-likelihood.

$$l'(\theta) = \frac{\partial}{\partial\theta} l_{uncen}(\theta) + \frac{\partial}{\partial\theta} l_{cen}(\theta), \tag{15}$$

where the derivatives of the uncensored and censored data are as given below.

$$\frac{\partial}{\partial\theta} l_{uncen}(\theta) = \sum_{uncen} z_i + \sum_{uncen} z_i \, exp(\theta z_i) \ln(S_0(t_i)), \tag{16}$$

$$\frac{\partial}{\partial\theta} l_{cen}(\theta) = \sum_{cen} \frac{\frac{\partial}{\partial\theta}(S(l_i|z_i) - S(r_i|z_i))}{S(l_i|z_i) - S(r_i|z_i)}$$

$$= \sum_{cen} \frac{z_i\, e^{\theta z_i} \ln\left[S_0\left(l_i\right)\right]\, \left[S_0\left(l_i\right)\right]^{e^{\theta z_i}} - z_i\, e^{\theta z_i} \ln\left[S_0\left(r_i\right)\right]\, \left[S_0\left(r_i\right)\right]^{e^{\theta z_i}}}{S\left(l_i|z_i\right) - S\left(r_i|z_i\right)}. \qquad (17)$$

This expression does not involve the baseline density, $f_0\left(t\right)$. While there is not a general, closed form solution to (15), it can be solved numerically.

We now have the framework necessary to set up an algorithm to find not only the maximum likelihood estimate of the regression parameter θ, but also estimate the baseline survival function, $S_0\left(t\right)$, in a Cox PH model where the distribution of covariate values has mass at 0. Under these conditions, the algorithm is as follows:

1. Estimate $S_0^{(1)}(t)$ via SCE (or NPMLE) only using the data at baseline level. That is, use all observations for which $z_i = 0$.
2. Estimate $\theta^{(1)}$ by solving for the root of (15) using $S_0^{(1)}(t)$.
3. Find $\widetilde{t_i} = S_0^{(1)^{-1}}\left[S_0^{(1)}\left(t_i\right)^{exp\left(\theta^{(1)} z_i\right)}\right]$; details of this step are given in the paragraph below. One can find $\tilde{l_i}$ and $\tilde{r_i}$ in the same fashion. Note that if $z_i = 0$, then $\widetilde{t_i} = t_i$ by definition of the Cox PH model. This is the key step which provides the data required to estimate the baseline survival function more accurately using all of the data.
4. Estimate $S_0^{(2)}(t)$ via SCE (or NPMLE) using *all* of the $\widetilde{t_i}$, $\widetilde{l_i}$, & $\widetilde{r_i}$ as data.
5. Estimate $\theta^{(2)}$ by solving for the root of 15 and using $S_0^{(2)}(t)$ to solve for the necessary probabilities in it.
6. Repeat steps (3)-(5) until a convergence criterion is met. Throughout this paper, we use the convergence criterion of a difference of 0.0001 between successive estimates of θ.

We now provide a justification for step 3 of the above algorithm. By the probability integral transformation, if we define $u_i = S_0^{(1)}\left(t_i\right)^{exp\left(\theta^{(1)} z_i\right)}$, then the $u_i's$ have a uniform distribution. To scale these back to the baseline survival function, we need to find $\widetilde{t_i} = \inf_t \left\{S_0^{(1)}(t) \leq u_i\right\}$. The correct choice is $\widetilde{t_i} = S_0^{(1)^{-1}}\left(u_i\right) = S_0^{(1)^{-1}}\left[S_0^{(1)}\left(t_i\right)^{exp\left(\theta^{(1)} z_i\right)}\right]$.

To illustrate the usefulness of this algorithm, a simulation study was done. In all cases, the baseline density, $f_0\left(t\right)$, was taken to be exponential with mean 10. The censoring mechanism consists of intervals whose left end points as well as lengths are independent exponentially distributed random variables. The covariates, Z_i, were generated from a $Binomial\left(1,\, 0.5\right)$ distribution where the true covariate effect is $\theta = 1$. As a graphical aide, graphs of the final estimate of $F_0\left(t\right)$ with sample sizes $n = 250$ and 1000 are given in Figure 1. In each of these figures, the empirical CDF (ECDF) of the uncensored data is given as a starting point, the true CDF, and the fitted CDF are given. In each case, the fitted CDF moves away from the original ECDF towards the true CDF. As the sample size, n, increases, the estimated CDF does a significantly better job at estimating the true distribution.

To illustrate the estimation of the covariate effect θ, $N = 100$ samples were simulated with sample sizes $n = 100$, 250, 500, and 1000. Again, the true value

Figure 1 – Cox PH model, with Discrete Covariates

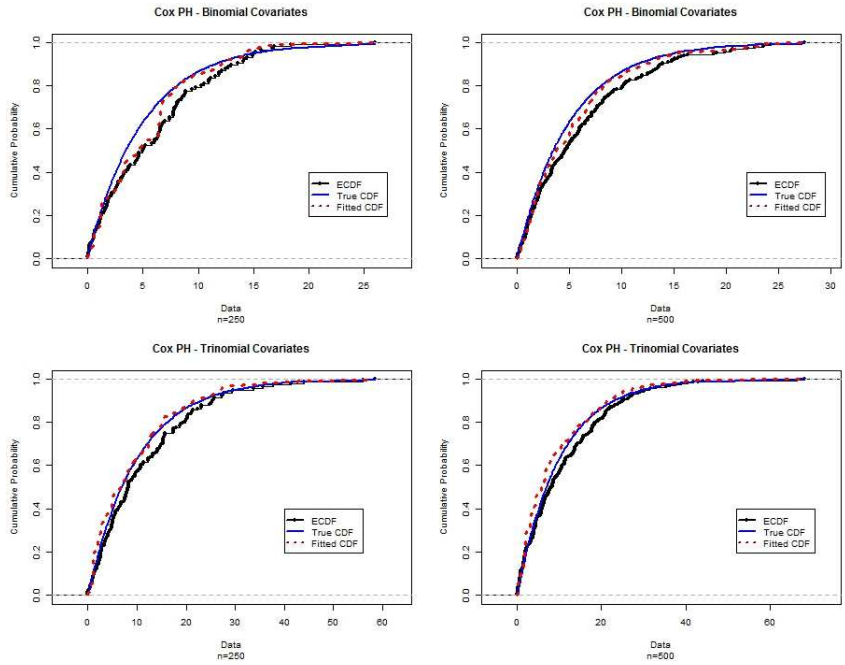

TABLE 1
Simulations for Cox PH model, Binomial Covariates

	$\hat{\theta}_{MLE}$	$MSE(\theta)$	$\bar{p}$	$S_{\bar{p}}$
n=100	0.9276	0.0697	0.2343	0.0384
n=250	0.9806	0.0231	0.2344	0.0302
n=500	0.9922	0.0083	0.2356	0.0191
n=1000	0.9948	0.0056	0.2326	0.0126

of θ was set equal to 1. The results of this simulation study are given in Table 1. Here, we see that as the sample size increases, the accuracy of $\hat{\theta}_{MLE}$ increases. Additionally, it's variability decreases, as is seen by inspecting the MSE in the table. Finally, $\bar{p}$ represents the average amount of censoring in $N = 100$ samples, and $S_{\bar{p}}$ is the standard deviation of the amount of censoring. On the whole, roughly 23% of these observations are censored. Even under this relatively high amount of censoring, the fitted CDF and MLE of θ perform remarkably well. In the above analysis the covariate can only take on two values, $Z = 0$ or 1. Next we consider covariates coming from a trinomial distribution. In all cases, the baseline density, $f_0(t)$, was taken to be exponential with mean 10. The censoring mechanism is the same as above. The covariates, Z_i, were generated from a $Trinomial\,(0.6,\, 0.2,\, 0.2)$ distribution and the true covariate effect is $\theta = 1$. Plots of this procedure are given in Figure 1. As with Binomial covariates, the estimate of the baseline distribution is quite good and improves as the sample size increases.

TABLE 2
Simulations for Cox PH model, Trinomial Covariates

	$\hat{\theta}_{MLE}$	$MSE(\theta)$	$\bar{p}$	$S_{\bar{p}}$
n=250	0.8708	0.0648	0.2118	0.0309
n=500	0.9318	0.0358	0.2104	0.0357
n=1000	0.9683	0.0085	0.2146	0.122

Next, multiple simulations were run to show the effectiveness of this methodology with respect to the trinomial regression parameter, θ. Samples of size $n = 250$, 500, and 1000 were considered, and $N = 100$ replications of each sample were studied. The results of this simulation study are given in Table 2. As in the case of $Binomial\,(1,\,0.5)$ covariates, the estimate of θ becomes more accurate as the sample size increases. Generally, samples of size $n = 500$ were stable, with samples of size $n = 1000$ giving consistently accurate results. In all simulations, roughly 21% of these observations are censored. Even under this relatively high amount of censoring, the fitted CDF and MLE of θ perform remarkably well. The authors have used values of θ ranging from 1 to 100 in simulations, along with Gamma and Weibull distriubtions. In all cases, the algorithm performed similarly to the results presented in this paper.

3. Accelerated Failure Time Model

We now consider the accelerated failure time model in the presence of middle censoring. Recall that the accelerated failure time (AFT) model has survival function of the form

$$S\,(t|Z) = S_0\left(t\,e^{\theta Z}\right) \tag{18}$$

Equivalently, the density function is given by

$$f\,(t|Z) = -\frac{\partial}{\partial t} S\,(t|Z) = e^{\theta Z}\,f_0\left(t\,e^{\theta Z}\right) \tag{19}$$

This model has been used for right censored and interval censored data. It is common practice in engineering and in reliability studies to assume that the baseline survival function, $S_0(t)$, is known. One of the more commonly assumed lifetime distributions is the exponential distribution. Thus our procedure for estimating the baseline survival function will provide a basis for selecting such parametric forms.

We first recall the fact that the Cox PH assumption is equivalent to the AFT assumption when the baseline distribution is an exponential distribution. To see this more clearly, if $T \sim Exp(a)$ then

$$S(t) = e^{-a\,t} \quad for\, t > 0 \tag{20}$$

Hence

$$S_0\left(t\,e^{\theta Z}\right) = \exp\left[-a\,t\,e^{\theta Z}\right] = \left(\exp\left[-a\,t\right]\right)^{e^{\theta Z}} = S_0(t)^{e^{\theta Z}} \tag{21}$$

Thus these models are indeed equivalent. If one assumes that the true distribution of lifetimes is an exponential distribution, then the same methodology described in Section 2 for the Cox PH model can be used. To make this explicit, the algorithm is as follows:

1. Estimate $S_0^{(1)}(t)$ via SCE (or NPMLE) using all of the data.
2. Estimate $\theta^{(1)}$ by solving for the root of 15 and using $S_0^{(1)}(t)$ to solve for the necessary probabilities in it.
3. Find $\widetilde{t_i} = t_i \exp\left(\theta^{(1)} z_i\right)$; details of this step are given in the paragraph below. One can find $\tilde{l_i}$ and $\tilde{r_i}$ in the same fashion.
4. Estimate $S_0^{(2)}(t)$ via SCE (or NPMLE) using all of the $\widetilde{t_i}$, $\widetilde{l_i}$, & $\widetilde{r_i}$ as your data.
5. Estimate $\theta^{(2)}$ by solving for the root of 15 and using $S_0^{(2)}(t)$ to solve for the necessary probabilities in it.
6. Repeat steps (3)-(5) until convergence criteria is met.

Again to see why step 3 in the above algorithm is correct, define the following

$$u_i = S_0^{(1)}\left(t_i \exp\left(\theta^{(1)} z_i\right)\right). \tag{22}$$

Then the $u_i's$ have a Uniform distribution by the probability integral transformation theorem. To scale these back to the baseline survival function, we need to find $\widetilde{t_i} = \inf_t \left\{S_0^{(1)}(t) \leq u_i\right\}$. The correct choice is

$$\widetilde{t_i} = S_0(u_i) = S_0^{(1)^{-1}}\left[S_0^{(1)}\left(t_i \exp\left(\theta^{(1)} z_i\right)\right)\right] = t_i \exp\left(\theta^{(1)} z_i\right) \tag{23}$$

A major addition to modelling the data in this manner is that one can check if the true baseline density is exponential or not. This is accomplished with a standard goodness of fit test, such as the Kolmogorov-Smirnov test. A program was written to study the semi-parametric AFT model. We considered the case where the baseline density, $f_0(t)$, is exponential with mean 10. The censoring mechanism is the same as in Section 2. The covariates, Z_i, were generated from a $Binomial(1, 0.5)$ distribution and the true covariate effect is $\theta = 1$. As a graphical aide, graphs of the final estimate of $F_0(t)$ with sample sizes $n = 100$, 250, and 500 are given in Figure 2. In each of these figures, the empirical CDF of the uncensored data is given as a starting point, the true CDF, and the fitted CDF are also given. In each case, the fitted CDF moves away from the original ECDF towards the true CDF. Again, as the sample size, n, increases, the estimated CDF does a significantly better job at estimating the true distribution. To estimate the covariate effect, θ, $N = 100$ samples were simulated with size $n = 100$, 250, and 500. Again, the true value of θ was chosen to be 1. The results of this simulation study are given in Table 3. Here, we see that as the sample size increases, the accuracy of $\hat{\theta}_{MLE}$ increases. Additionally, its variability decreases, as is seen by inspecting the MSE in the table. Finally, $\bar{p}$ represents the average amount of censoring in these $N = 100$ samples, and $S_{\bar{p}}$ is the standard deviation of the amount of censoring. With a high amount of censoring, 23%, this procedure gives extremely accurate results. Under this semiparametric framework, we are not confined to covariates coming from a Multinomial distribution. To illustrate this, we considered the same set-up as above, except now the covariates, Z_i, were generated from an $Exponential(1)$

Figure 2 – Exponential AFT model

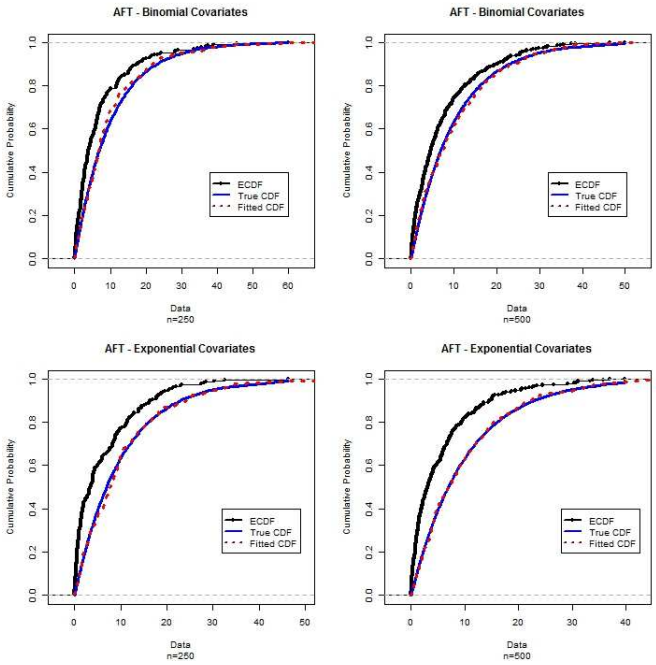

TABLE 3
Simulations for Exponential AFT model, Binomial Covariates

	$\hat{\theta}_{MLE}$	$MSE\,(\theta)$	$\bar{p}$	$S_{\bar{p}}$
n=100	1.0396	0.0752	0.2422	0.0407
n=250	1.0214	0.0278	0.2291	0.0271
n=500	0.9996	0.0156	0.2351	0.0178

TABLE 4
Simulations for Exponential AFT model, Exponential Covariates

	$\hat{\theta}_{MLE}$	$MSE(\theta)$	$\bar{p}$	$S_{\bar{p}}$
n=100	0.9911	0.0445	0.2198	0.0439
n=250	0.9922	0.0202	0.2163	0.0236
n=500	0.9943	0.0067	0.2163	0.0179

distribution with a true covariate effect of $\theta = 1$. Graphs of the ECDF and final estimate of $F_0(t)$ with sample sizes $n = 100$, 250, and 500 are given in Figure 2. Again, the true CDF is very accurately estimated.

As before, $N = 100$ samples were simulated with size $n = 100$, 250, and 500 with a true value of $\theta = 1$. The results of this simulation study are given in Table 4. Here, we see that as the sample size increases, the accuracy of $\hat{\theta}_{MLE}$ increases. Additionally, its variability decreases, as is seen by inspecting the MSE in the table. With a high amount of censoring, roughly 21%, this procedure yields extremely accurate results. The authors have used values of θ ranging from 1 to 100 in simulations, along with Gamma and Weibull distriubtions. In all cases, the algorithm performed similarly to the results presented in this paper. Note that for the time being, we have only dealt with one covariate; we believe that this approach will work when dealing with multiple covariates, although this needs to be studied further.

4. Application to Contingent Valuation

One major area of research in economics is Cost Benefit Analysis (CBA). The goal of CBA is to decide which policy maximizes the welfare of the population. In environmental economics, this valuation is obtained through Contingent Valuation Methods (CVM). In CVM surveys, people are asked how much they value a certain natural resource. This is done in two different ways, indirectly and directly (cf. Hanley, Shogren and White (2001)). The indirect valuation method uses indirect measures to reveal how much a natural resource is worth. The direct method of valuation is much more straightforward, as is suggested by its name. With these types of surveys, people are asked about their Willingness To Pay (WTP) to increase the quality of a natural resource. We refer the reader to Hanley, Shogren and White (2001), Braden & Kolstad (1991), and Smith (1993) for further details on CVM and WTP.

4.1. Scandinavian WTP Example

The following data was obtained by Cecilia Hakansson from Sweden and Katja Parkkila from Finland in a 2004 survey on people's WTP (see Hakansson (2007)). This survey was a study in which respondents were asked a classic and interval open ended (CIOE) questions in order to investigate the economic feasibility of altering the path of a river around a major hydro-power plant on the Vindel River, in Northern Sweden near the Finnish border. The proposed plan would reduce the production of electricity by allocating more water to flow through the waterways built to let the wild salmon swim upstream to spawn, and hence

TABLE 5
Averages for WTP Data by Country

	Left Data	Exact Data	Right Data	Fitted CDF
Swedes	17.31	30.91	42.40	31.11
Finns	20.65	30.36	45.28	30.56

this would increase the overall number of wild salmon in the river. The goal of this survey was to learn about citizen's WTP to increase the volume of salmon reaching the spawning area every year.

The proportion of respondents choosing to give an interval for their WTP was roughly 50%, a very high level of censored data. Another interesting observation in the data is that over 95% of WTP values given were rounded to the nearest 5 Swedish kronor (SEK) increment, such as 20, 50, or 100 SEK. Due to this tendancy of people rounding monetary values to the nearest 5 or 10, interval (censored) data can be much more useful in trying to estimate the true average WTP of a population.

Additionally, we computed the average values for all of the left and right endpoints of the interval data, and the average value of the exact data separately for each country. This was done in order to compare them to the theoretical average value from the fitted distribution. See Table 5 for the values. The mean of the left interval data is less than the mean of the exact data and both are less than the mean of the right interval data for both countries. Moreover, the average from the fitted CDF is very close to the average of exact data and is between the averages of the left and right interval data as well.

4.2. Data Analysis

A model with a single covariate was then fit to the Swedish data. The choice of the single covariate, β = annual income by incremental category, was determined by economic theory. When this data was fit, the effect of annual income was estimated to be 0.0747, with an associated p-value of 0.0060 (calculated using the asymptotic normality of MLE's with $n = 132$ observations in this dataset). This is consistent with the interpretation that people with higher incomes are willing to pay more to preserve natural resources.

This same model was fit for the Finnish data with the same choice of covariate, β = annual income by incremental category. Now the effect of annual income was estimated to be 0.0910, with an associated p-value ¡0.0001 (calculated the same as before but with $n = 203$ observations in this dataset). Thus in both Scandinavian countries, people with higher incomes are willing to pay more to help the salmon stock in their countries.

This survey study on WTP is an excellent real-world problem modeled by implementating a middle censoring algorithm. Previously, economists mostly dealt with only exact data or dichotomous choice questions due to complications in analyzing data. If one assumes a middle censored model, a much richer class of questions can be asked, resulting in much more meaningful data.

5. Conclusion

As shown in this paper, this approach does well to fit data when it is simulated from an Exponential distribution. While not presented here, this method also performs well when the underlying distribution is Gamma or Weibull. Further research needs to be done in other cases, such as when the data comes from a Log-Normal distribution. This method also does well when considering one covariate and when applied to a real world dataset.

A limitation in this paper is how continuous covariates are dealth with. Currently it is common practice to discretize them into categories. For example, blood pressure is a continuous variable but it is usually grouped into one of four broad categories, low, normal, elevated and high. In economics, people are often asked about their income, but these responses are usually binned into either groups by $\$10,000$ increments or into other groups, such as low income, lower middle-class, upper middle-class, wealthy, and ultra wealthy. The approach presented in this paper would apply when it makes sense to bin a continuous covariate into discrete groups. An approach to handle continuous covariates without categorizing them is an area requiring further reseach. Finally, we have only dealt with a single covariate at this point, but see no serious impediment to studying several covariates.

References

K. Arrow, R. Solow (1993). *Report of the noaa panel on contingent valuation.* Federal Register, 58, pp. 4601–14.

M. Ayer, H. Brunk, G. Ewing, W. Reid, E. Silverman (1955). *An empirical distribution function for sampling with incomplete information.* The Annals of Mathematical Statistics, 26, no. 4, pp. 641–647.

b. Efron (1977). *The two-sample problem with censored data.* In *Proc. Fifth Berkeley Symp. Math. Statist. Probab.* vol. 4, pp. 831–853.

N. Bennett, S. Iyer, S. Jammalamadaka (2017). *Analysis of gamma and weibull lifetime data under a general censoring scheme and in the presence of covariates.* Communications in Statistics - Theory and Methods, 46, no. 5, pp. 2277–2289.

J. Braden, C. Kolstad (1991). *Measuring the demand for environmental quality.* Contributions to Economic Analysis (Países Bajos), 198, pp. 333–355.

D. Cox (1972). *Regression models and life-tables.* Journal of the Royal Statistical Society B, 34, pp. 187–220.

C. Hakansson (2007). *Cost-Benefit Analysis and Valuation Uncertainty.* Ph.D. thesis, Acta Universitatis Agreculturae Sueciae.

N. Hanley, J. Shogren, B. White, E. P. (Firm) (2001). *Introduction to environmental economics.* Oxford University Press.

S. Iyer, S. Jammalamadaka, D. Kundu (2008). *Analysis of middle-censored data with exponential lifetime distributions*. Journal of Statistical Planning and Inference, 138, no. 11, pp. 3550–3560.

S. Jammalamadaka, S. Iyer (2004). *Approximate self consistency for middle-censored data*. Journal of Statistical Planning and Inference, 124, no. 1, pp. 75–86.

S. Jammalamadaka, V. Mangalam (2003). *Nonparametric estimation for middle-censored data*. Journal of Nonparametric Statistics, 15, no. 2, pp. 253–265.

S. Jammalamadaka, V. Mangalam (2009). *A general censoring scheme for circular data*. Statistical Methodology, 6, no. 3, pp. 280–289.

S. Jammalamadaka, S. Prasad, P. Sankaran (2016). *A semi-parametric regression model for analysis of middle censored lifetime data*. Statistica, 76, no. 1, pp. 27–40.

E. Kaplan, P. Meier (1958). *Nonparametric estimation from incomplete observations*. Journal of the American Statistical Association, 53, pp. 457–481.

P. Shen (2011). *The nonparametric maximum likelihood estimator for middle-censored data*. Journal of Statistical Planning and Inference, 141, pp. 2494–2499.

V. Smith (1993). *Nonmarket valuation of environmental resources: an interpretive appraisal*. Land Economics, 69, no. 1, pp. 1–26.

B. Turnbull (1976). *The empirical distribution function with arbitrarily grouped, censored and truncated data*. Journal of the Royal Statistical Society B, 38, pp. 290–295.

J. Wellner (1982). *Asymptotic optimality of the product limit estimator*. The Annals of Statistics, 10, pp. 595–602.

Summary

We are interested in estimating the distribution of lifetimes, also called survival times, subject to a general censoring scheme called "middle censoring" (see Jammalamadaka and Mangalam (2003)). Both the Cox proportional hazards (Cox PH) and accelerated failure time (AFT) models are considered since each model has a baseline distribution function that is modified by the presence of covariates. The key contribution presented is the estimation of the effect of the covariate as well as the baseline distribution function. We conclude with an application to a contingent valuation study.

Keywords: Middle censoring; Cox PH model; Accelerated failure time model (AFT); Contingent valuation; Willingness to pay.