International Journal of Applied Science and Engineering
2018. 15, 1: 47-57

Prediction for a Future Number of Failures Based on Right-Censored Data with Indeterminate Survival Times and Censoring Status

B. L. Lee[a*], Steven M. Crunk[a], Maheen Khan[b] and William B. Fairley[c]

[a]*Department of Mathematics and Statistics, San José State University, San José, CA, USA*
[b]*Cengage Learning, San Francisco, CA, USA*
[c]*Analysis & Inference, Inc., Springfield, PA, USA*

Abstract: We consider parametric inference for right-censored data where the survival times and the corresponding censoring status cannot be determined from the data. We develop point and interval prediction for a future number of surviving units. The performance of the proposed methods are assessed by means of a simulation study, and an application of the proposed methods is illustrated using a real data set.

Keywords: Right-censored data; parametric; survival distribution; loss function; prediction interval

1. Introduction

Survival data are typically not fully observed. A common form of incompleteness is right censoring, where the time to an event is only known to exceed some value. For example, in reliability studies, a right-censored survival time arises if a unit placed on test is still in operation at the end of the study. When the survival times and the corresponding censoring status are fully determined from the data, the survival function is traditionally estimated using the Kaplan-Meier nonparametric estimator [1].

In this paper, we consider the analysis of right-censored data with severe incompleteness, a real data scenario from a consulting project, described as follows. A batch of medical devices was manufactured, distributed and put into service on different dates. Subsequently, the dates of failure were reported to the manufacturer from various clients. Due to confidentiality issues, the reports did not include the serial numbers of the devices, so that none of the dates of failure could be matched to the complete list of dates that the devices began service (a matching installation and failure date could be traced back to a patient). In other words, neither the survival times nor the censoring status could be determined from the given data; all that could be determined were the number of units that began service and were in service on each day of the study. Of interest is the prediction of the number of future failures, so as to decide whether a recall is warranted for the remaining units in service.

To our knowledge, the existing literature does not provide guidance on the analysis of such data. The focus of this paper is to develop a method for estimating a parametric survival distribution assumed for the data described in the preceding paragraph, and to construct a prediction interval for the number of units in service on a specified day beyond the end of the study. While parametric

*Corresponding author; e-mail: beeleng.lee@sjsu.edu
doi: 10.6703/IJASE.201802_15(1).047
©2018 Chaoyang University of Technology, 1727-2394
Received 31 July 2017
Revised 12 January 2018
Accepted 18 January 2018

modeling has long been criticized for imposing distributional assumptions on the data, it is a point of departure for analyzing nontraditional data and a reference for future research. The rest of this paper is organized as follows. Section 2 develops the proposed method, the performance of which is assessed in section 3 by means of a simulation study. Section 4 illustrates the application of the proposed method with a real data set. Finally, some concluding remarks based on the obtained results of simulations are given in section 5.

2. Methods

Let $N_B(l)$ and $N_I(l)$ denote the number of units that begin service and are in service, respectively, on day l. Here $l = 1, 2, ..., l_B, ..., l_E, ...$, where $l = 1$ corresponds to the earliest date that one or more units begin service, $l = l_B$ corresponds to the date that the last unit begins service, and l_E corresponds to the end date of the study. Furthermore, let $N_I(k,l)$ denote the number of units that are in service on day l among those which begin service on day k, where $k \leq l$; we have $N_I(l) = \sum_{k=1}^{l} N_I(k,l)$. The quantity $N_I(k,l)$ is unobserved for all k and l and is introduced for the purpose of studying the properties of $N_I(l)$. The observations consist of realizations of $N_B(l)$ and $N_I(l)$ for $l = 1, ..., l_E$, which we denote by $n_B(l)$ and $n_I(l)$, respectively. Figure 1 illustrates our notation. Of interest is the prediction of $N_I(l)$ for $l > l_E$ at the end of the study, which we develop as follows.

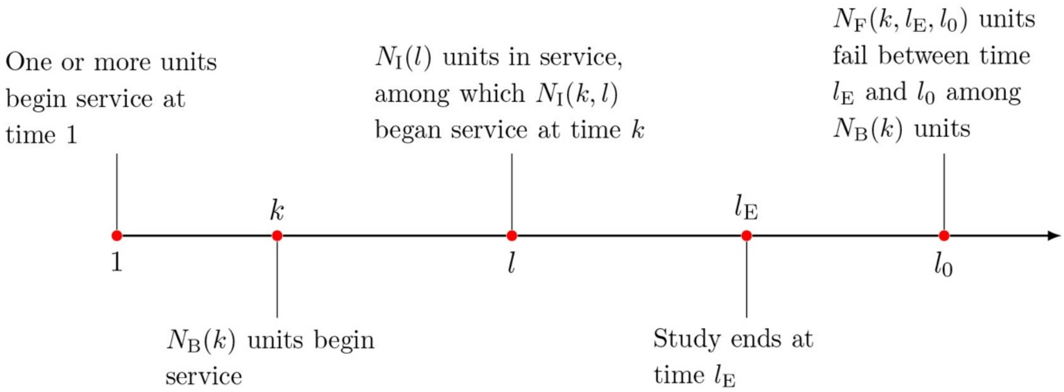

Figure 1. An illustration of notation.

2.1. Estimation of survival function

We assume that the survival times of all the units are independent and identically distributed with survival function S_θ, where θ is a possibly multidimensional unknown parameter. For each of the $N_B(k)$ units which begins service on day k, the probability of the unit being in service on day l, where $k \leq l$, is given by $S_\theta(l - k)$. It follows that, given $N_B(k) = n_B(k)$, $N_I(k,l)$ has a binomial distribution with a total number of trials equal to $n_B(k)$ and a success probability of $S_\theta(l - k)$. Thus, for $l = 1, ..., l_E$,

$$\begin{aligned} E[N_I(l)|n_B(1), \dots, n_B(l_E)] &= E\left[\sum_{k=1}^{l} N_I(k,l)|n_B(1), \dots, n_B(l_E)\right] \\ &= \sum_{k=1}^{l} E[N_I(k,l)|n_B(k)] \\ &= \sum_{k=1}^{l} n_B(k) S_\theta(l-k). \end{aligned} \tag{1}$$

To estimate θ, we minimize

$$\sum_{l=1}^{l_{\mathrm{E}}} g\Big(n_{\mathrm{I}}(l) - \sum_{k=1}^{l} n_{\mathrm{B}}(k) S_\theta(l-k)\Big) \tag{2}$$

with respect to θ, where g is a specified loss function. For example, $g(x) = x^2$ corresponds to a squared error loss, and $g(x) = |x|$ corresponds to an absolute error loss. The summands in (2) are not independent of one another since, for instance, $n_{\mathrm{I}}(l)$depends on $n_{\mathrm{I}}(l-1)$.

2.2. Prediction of number of units in service

2.2.1. Point prediction

$$\mathrm{E}[N_{\mathrm{I}}(l) | n_{\mathrm{B}}(1), \ldots, n_{\mathrm{B}}(l_{\mathrm{E}})]$$

Let $\hat{\theta}$ denote the estimate of θ obtained by minimizing the loss function (2), and let $l_0 > l_{\mathrm{E}}$. A naive method for predicting the value of $N_{\mathrm{I}}(l_0)$ is obtained by replacing S_θ with $S_{\hat{\theta}}$ in (1):

$$\tilde{n}_{\mathrm{I}}(l_0) = \sum_{k=1}^{l_0} n_{\mathrm{B}}(k) S_{\hat{\theta}}(l_0-k) = \sum_{k=1}^{l_{\mathrm{B}}} n_{\mathrm{B}}(k) S_{\hat{\theta}}(l_0-k),$$

where the second equality follows from the fact that $n_{\mathrm{B}}(k) = 0$ for $k > l_{\mathrm{B}}$. This naive method suffers from the defect that $\tilde{n}_{\mathrm{I}}(l_0)$ is not necessarily less than or equal to $n_{\mathrm{I}}(l_{\mathrm{E}})$, which led to the consideration of an alternative method. Let $N_{\mathrm{F}}(k, l_{\mathrm{E}}, l_0)$ denote the number of units that fail between day l_{E} and day l_0 among those that begin service on day k, where $k \le l_{\mathrm{E}}$. By definition,

$$N_{\mathrm{I}}(l_0) = N_{\mathrm{I}}(l_{\mathrm{E}}) - \sum_{k=1}^{l_{\mathrm{B}}} N_{\mathrm{F}}(k, l_{\mathrm{E}}, l_0) \tag{3}$$

The quantity $N_{\mathrm{F}}(k, l_{\mathrm{E}}, l_0)$ is unobserved for all k but given $N_{\mathrm{I}}(k, l_{\mathrm{E}}) = n_{\mathrm{I}}(k, l_{\mathrm{E}})$, it has a binomial distribution with a total number of trials $n_{\mathrm{I}}(k, l_{\mathrm{E}})$ and a success probability of

$$p(k, l_0) = \frac{S_\theta(l_{\mathrm{E}}-k) - S_\theta(l_0-k)}{S_\theta(l_{\mathrm{E}}-k)}$$

It follows that

$$\mathrm{E}[N_{\mathrm{F}}(k, l_{\mathrm{E}}, l_0) | N_{\mathrm{I}}(k, l_{\mathrm{E}})] = N_{\mathrm{I}}(k, l_{\mathrm{E}}) p(k, l_0). \tag{4}$$

The proposed predicted value of $N_{\mathrm{I}}(l_0)$, denoted $\hat{n}_{\mathrm{I}}(l_0)$, is obtained by replacing the unobserved $N_{\mathrm{F}}(k, l_{\mathrm{E}}, l_0)$ in (3) with an estimate of $\mathrm{E}[N_{\mathrm{F}}(k, l_{\mathrm{E}}, l_0) | N_{\mathrm{I}}(k, l_{\mathrm{E}})]$ given in (4). The unknown quantities $N_{\mathrm{I}}(k, l_{\mathrm{E}})$ and $p(k, l_0)$ in (4) are estimated by

$$\hat{n}_{\mathrm{I}}(k, l_E) = n_B(k) S_{\hat{\theta}}(l_E - k)$$

and

$$\hat{p}(k, l_0) = \frac{S_{\hat{\theta}}(l_{\mathrm{E}}-k) - S_{\hat{\theta}}(l_0-k)}{S_{\hat{\theta}}(l_{\mathrm{E}}-k)}$$

respectively. That is,

$$\begin{aligned} \hat{n}_{\mathrm{I}}(l_0) &= n_{\mathrm{I}}(l_{\mathrm{E}}) - \sum_{k=1}^{l_{\mathrm{B}}} \hat{n}_{\mathrm{I}}(k, l_{\mathrm{E}}) \hat{p}(k, l_0) \\ &= n_{\mathrm{I}}(l_{\mathrm{E}}) - \sum_{k=1}^{l_{\mathrm{B}}} n_{\mathrm{B}}(k) [S_{\hat{\theta}}(l_{\mathrm{E}}-k) - S_{\hat{\theta}}(l_0-k)]. \end{aligned}$$

2.2.2. Interval prediction

We use the normal approximation to compute a $100(1-\alpha)\%$ conditional prediction interval for $N_I(l_0)$, conditional on the observed data $\{(n_B(l), n_I(l)) : l = 1,...,l_E\}$:

$$n_I(l_0) \pm z_{\alpha/2}\sqrt{\mathrm{var}(\epsilon_0|\mathrm{data})}$$

where $z_{\alpha/2}$ denotes the upper $\alpha/2$ quantile of the standard normal distribution and $_0$ denotes the prediction error. To approximate $\mathrm{var}(\epsilon_0|\mathrm{data})$, we begin by considering the case where the parameter θ is known:

$$\begin{aligned}\epsilon_0 &= \widehat{N}_I(l_0) - N_I(l_0)\\ &= \left\{N_I(l_E) - \sum_{k=1}^{l_B} N_B(k)[S_\theta(l_E-k) - S_\theta(l_0-k)]\right\} - \left[N_I(l_E) - \sum_{k=1}^{l_E} N_F(k, l_E, l_0)\right]\\ &= \sum_{k=1}^{l_E} N_F(k, l_E, l_0) - \sum_{k=1}^{l_B} N_B(k)[S_\theta(l_E-k) - S_\theta(l_0-k)].\end{aligned}$$

In this case,

$$\begin{aligned}\mathrm{var}(\epsilon_0|\mathrm{data}) &= \mathrm{var}\left(\sum_{k=1}^{l_E} N_F(k, l_E, l_0)|\mathrm{data}\right)\\ &= \sum_{k=1}^{l_E} \mathrm{var}\left(N_F(k, l_E, l_0)|n_I(k, l_E)\right)\\ &= \sum_{k=1}^{l_E} n_I(k, l_E)p(k, l_0)[1-p(k, l_0)].\end{aligned}$$

In practice, the parameter θ is unknown and must be estimated from the observed data; in this case, the prediction error variance is intractable to compute. We adopt a common approach in time series analysis [2–4], which is to approximate the prediction error variance by disregarding the variability introduced by the estimation of θ. That is, we approximate $\mathrm{var}(\epsilon_0 \mid \mathrm{data})$ with

$$\begin{aligned}\widehat{var}(\epsilon_0|\mathrm{data}) &= \sum_{k=1}^{l_E} \hat{n}_I(k, l_E)\hat{p}(k, l_0)[1-\hat{p}(k, l_0)]\\ &= n_B(k)[S_\theta(l_E-k) - S_\theta(l_0-k)]\left[\frac{S_\theta(l_0-k)}{S_\theta(l_E-k)}\right]\\ &= \sum_{k=1}^{l_B} n_B(k)[S_\theta(l_E-k) - S_\theta(l_0-k)]\left[\frac{S_\theta(l_0-k)}{S_\theta(l_E-k)}\right],\end{aligned} \tag{5}$$

where once again the final equality follows from the fact that $n_B(k) = 0$ for $k > l_B$. According to [5], in the context of time series forecasting, the effect of parameter uncertainty on the coverage of prediction intervals could be non-trivial for sample sizes smaller than about 50; however, such effect diminishes as the sample size gets larger. Although it is possible that resampling methods can improve the estimation of prediction variance, the application of these methods to the data considered here is non-trivial and is a topic for future research

3. Simulation study

To assess the performance of the methods described in section 2, we conducted a simulation study using the R statistical software version 3.2.2.

3.1. Design

The starting point for simulating data was to generate the underlying survival times. We considered two sample sizes, 50 and 100 units, and three commonly used distributions for survival data: Weibull, lognormal, and gamma. The Weibull and gamma distributions are characterized by a shape parameter and a scale parameter, while the lognormal distribution is characterized by a location parameter and a scale parameter; we denote the parameters by θ_1 (shape or location) and θ_2 (scale). The parameter values for each of the three distributions, shown in the first column of Table 1 and Table 2, were chosen such that the mean and standard deviation of the survival times are approximately 160 days and 60 days, respectively. Figure 2 shows a plot of the density functions of the survival distributions considered in our simulation study.

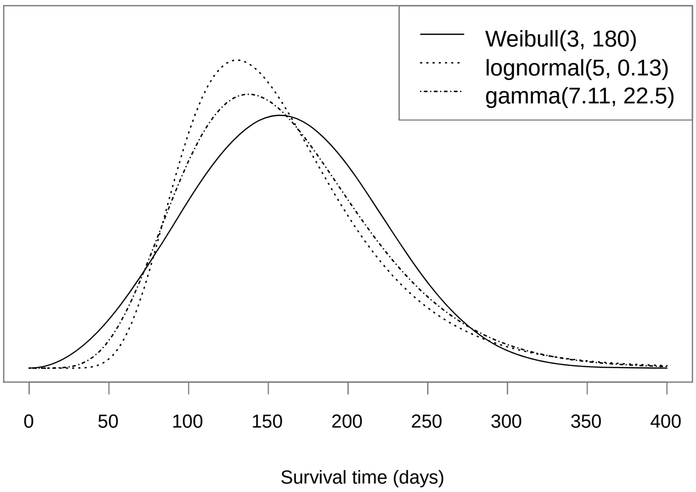

Figure 2. Density functions of the survival distributions considered in the simulation study.

To simulate the staggered entry of the units for each sample, we generated the days on which a unit begins service according to a Poisson distribution with mean 400. The days were subsequently shifted so that the first unit begins service on day one. If we denote the survival time of a unit by t and the day on which the unit begins service by d_B, the day on which the unit fails is calculated as $\lceil d_B + t \rceil$, where $\lceil \cdot \rceil$ denotes the ceiling function. We considered three fixed censoring times: l_E = 180, 240, and 300 days. The observable data, $n_B(l)$ and $n_I(l)$ for $l = 1,...,l_E$, were then calculated accordingly.

With three survival distributions, three censoring times, and two sample sizes, there is a total of 18 scenarios. For each scenario, 5000 data sets were generated, and the method described in section 2.1 was applied to each data set to estimate $\theta = (\theta_1, \theta_2)$ under no model misspecification. We considered both the squared and absolute error loss functions, which were minimized using the R function `nlminb` with default options, except that the relative tolerance was set to 0.0001. If convergence was not achieved in the minimization, the estimates were not included in the calculation of the summary statistics. For each scenario, we calculated the average percentage of censoring, the average of the parameter estimates, and the Monte Carlo standard errors of the parameter estimates across the 5000 data sets.

Additionally, we constructed 0.95 prediction intervals according to the method described in section 2.2, using the squared-error loss function and for each of the three survival distributions. The censoring time is 240 days and l_0 ranges from 241 to 300 days. We calculated the average observed and predicted number of units in service for each l_0, denoted $\bar{n}_I(l_0)$ and $\bar{\hat{n}}_I(l_0)$ respectively, where the average is taken over the 5000 simulations. We also computed the empirical coverage probabilities of the prediction intervals for each l_0.

3.2. Results

Table 1 and Table 2 present the results of parameter estimation under squared error loss with a sample size of 50 and 100 units, respectively. For each survival distribution, the bias and standard errors of the parameter estimates decrease as the sample size increases, and when the level of censoring decreases. Also, the proposed method produced the best results when the underlying survival times are distributed as lognormal, even when the sample size is small (50 units) and the level of censoring is heavy (approximately 50%).

Table 1. Simulation results for parameter estimation under no model misspecification and squared error loss based on a sample size of 50.

Survival distribution	Censoring time	Ave. % censored*	% converged	$\hat{\theta}_1$		$\hat{\theta}_2$	
				Avg.	S.e.	Avg.	S.e.
Weibull	180 days	48	100	3.29	1.72	186.78	55.91
$\theta_1 = 3$	240 days	31	100	3.11	0.59	181.07	11.54
$\theta_2 = 180$	300 days	14	100	3.10	0.47	180.50	9.38
Longnormal	180 days	53	100	5.02	0.12	0.14	0.10
$\theta_1 = 5$	240 days	37	100	5.01	0.06	0.13	0.05
$\theta_2 = 0.13$	300 days	22	100	5.00	0.05	0.13	0.04
Gamma	180 days	52	100	9.15	7.54	25.34	19.79
$\theta_1 = 7.11$	240 days	36	100	7.88	2.75	22.86	8.18
$\theta_2 = 22.5$	300 days	19	100	7.73	2.23	22.44	6.37

* Rounded off to the nearest integer.

Table 2. Simulation results for parameter estimation under no model misspecification and squared error loss based on a sample size of 100.

Survival distribution	Censoring time	Ave. % censored*	% converged	$\hat{\theta}_1$		$\hat{\theta}_2$	
				Avg.	S.e.	Avg.	S.e.
Weibull	180 days	51	100	3.15	0.75	183.06	23.39
$\theta_1 = 3$	240 days	35	100	3.06	0.41	180.96	8.46
$\theta_2 = 180$	300 days	18	100	3.05	0.33	180.52	6.68
Longnormal	180 days	57	100	5.01	0.08	0.14	0.06
$\theta_1 = 5$	240 days	43	100	5.00	0.04	0.13	0.03
$\theta_2 = 0.13$	300 days	28	100	5.00	0.04	0.13	0.03
Gamma	180 days	55	100	8.00	3.86	24.13	11.38
$\theta_1 = 7.11$	240 days	41	100	7.48	1.76	22.73	5.68
$\theta_2 = 22.5$	300 days	26	100	7.41	1.46	22.51	4.51

* Rounded off to the nearest integer.

Table 3 and Table 4 present the results of parameter estimation under absolute error loss with a sample size of 50 and 100 units, respectively. In contrast to estimation under squared error loss, where convergence of the minimization algorithm was achieved in virtually all the simulations, the percentage of convergence under absolute error loss ranges from 64% to 98%. Nevertheless, the results are similar to those observed in Table 1 and Table 2, even when the estimates for which convergence was not achieved were included in the calculation of the summary statistics.

Table 3. Simulation results for parameter estimation under no model misspecification and absolute error loss based on a sample size of 50.

Survival distribution	Censoring time	Ave. % censored*	% converged	$\hat{\theta}_1$ Avg.	$\hat{\theta}_1$ S.e.	$\hat{\theta}_2$ Avg.	$\hat{\theta}_2$ S.e.
Weibull	180 days	48	84	3.37	2.62	187.59	126.33
$\theta_1 = 3$	240 days	31	96	3.13	0.60	181.02	11.77
Longnormal	180 days	53	70	5.03	0.16	0.14	0.11
$\theta_1 = 5$	240 days	37	69	5.00	0.06	0.13	0.05
$\theta_2 = 0.13$	300 days	22	70	5.00	0.05	0.13	0.04
Gamma	180 days	52	78	8.79	4.98	24.82	18.41
$\theta_1 = 7.11$	240 days	36	94	7.92	2.75	22.72	8.21
$\theta_2 = 22.5$	300 days	20	97	7.76	2.24	22.37	6.39

* Rounded off to the nearest integer.

Table 4. Simulation results for parameter estimation under no model misspecification and absolute error loss based on a sample size of 100.

Survival distribution	Censoring time	Ave. % censored*	% converged	$\hat{\theta}_1$ Avg.	$\hat{\theta}_1$ S.e.	$\hat{\theta}_2$ Avg.	$\hat{\theta}_2$ S.e.
Weibull	180 days	51	81	3.16	0.75	182.61	24.75
$\theta_1 = 3$	240 days	35	95	3.06	0.42	180.95	8.69
$\theta_2 = 180$	300 days	18	98	3.06	0.33	180.53	6.72
Longnormal	180 days	57	66	5.01	0.09	0.14	0.07
$\theta_1 = 5$	240 days	43	64	5.01	0.04	0.13	0.03
$\theta_2 = 0.13$	300 days	28	64	5.00	0.04	0.13	0.03
Gamma	180 days	55	77	7.99	3.36	23.94	11.68
$\theta_1 = 7.11$	240 days	41	92	7.51	1.76	22.66	5.75
$\theta_2 = 22.5$	300 days	26	97	7.42	1.46	22.49	4.54

* Rounded off to the nearest integer.

Figure 3 compares the average observed and predicted number of units in service across 5000 simulations when the underlying and fitted survival distributions are both Weibull. There is little bias associated with point predictions even at 60 days beyond the end of 240 days of study. Figure 4 shows a plot of the empirical coverage probabilities of the corresponding 0.95 prediction intervals against the prediction horizons for sample sizes 50 and 100. The empirical coverage probabilities are close to the nominal level when the horizon is within 10 days, and they decline steadily as the horizon lengthens. The results for the lognormal and gamma distributions are very similar and we summarize the comparisons for all three distributions in Table 5.

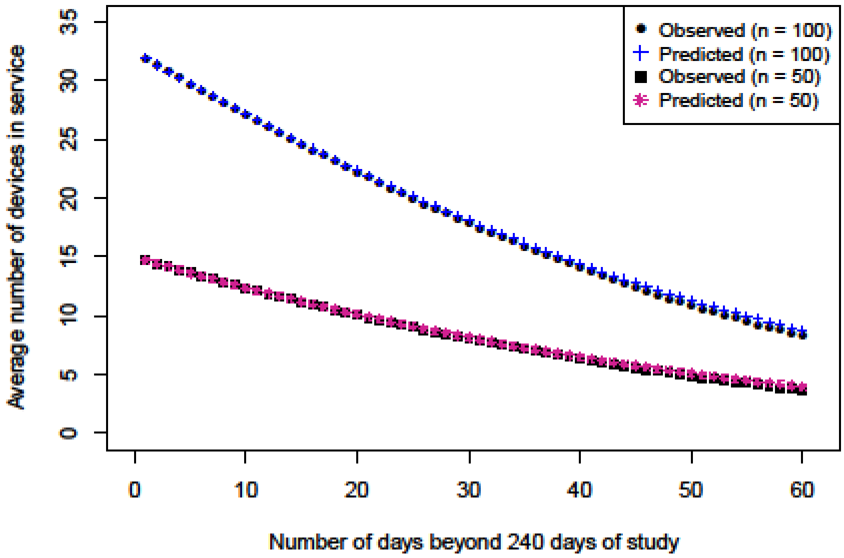

Figure 3. Average observed and predicted number of devices in service across 5000 simulations. The underlying and fitted survival distributions are Weibull.

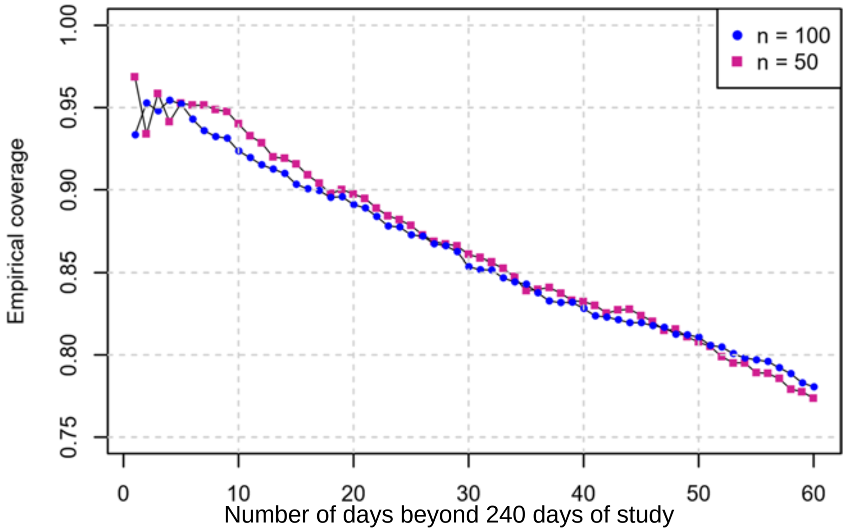

Figure 4. Empirical coverage probability of 0.95 prediction intervals based on 5000 simulations. The underlying and fitted survival distributions are Weibull.

4. Case Study

The impetus for the development of the method described in this paper was a data set provided to us by a medical device manufacturing firm. A collection of devices were manufactured together and distributed to providers. When a device was installed the provider would report the date back to the manufacturer. When a device failed the date of failure was likewise reported to the manufacturer. However, neither report would include any further identifying information so the survival time of any particular device was unknown. The only information available was the number of units installed at a given time and the number of units which had failed at a given time (and thus the number of units in service at any given time can be calculated). Rather than daily increments as described previously, the data here were provided as monthly aggregates.

Table 5. Simulation results for prediction of the number of units in service under squared error loss and censoring at 240 days.

		Sample Size					
		50 units			100 units		
Survival distribution	l_0	$\bar{n}_I(l_0)$	$\bar{\bar{n}}_I(l_0)$	Coverage	$\bar{n}_I(l_0)$	$\bar{\bar{n}}_I(l_0)$	Coverage
Weibull	270	8.1	8.2	0.86	17.8	18.1	0.85
$\theta_1 = 3, \theta_2 = 180$	300	3.7	4.0	0.77	8.3	8.7	0.78
Lognormal	270	7.0	7.0	0.90	15.2	15.3	0.90
$\theta_1 = 5, \theta_2 = 0.13$	300	3.9	4.0	0.83	8.3	8.5	0.84
Gamma	270	7.8	7.8	0.89	17.0	17.0	0.90
$\theta_1 = 7.11, \theta_2 = 22.5$	300	4.1	4.3	0.82	9.0	9.2	0.85

A single lot of 504 devices were manufactured. According to the convention described above the first month in which a device was installed is labeled month one. The last of the 504 devices was installed in month 23. Meanwhile, the first failure occurred in month nine and the data were reported to us for analysis at the end of month 36, by which time 206 devices had failed.

An attempt was made to fit each of a Weibull, lognormal and gamma distribution to the data using the methods described in section 2. The lognormal and gamma distributions provided similar results, and the value of the squared error loss function of (2) are 2786 and 8566 respectively. In the neighborhood of the parameter θ = (shape, scale) at which we might expect to minimize the loss function for the Weibull distribution (based on an assumption that we might expect similar results in terms of mean, variance, mode, skewness or other characteristics of this distribution as compared to the fitted lognormal or gamma distributions), the loss function is flat leaving no clear indication as to appropriate values of the parameters. Alternatively, allowing an initial shape parameter value of one, the default in R, or any combination of initial values in the border region of the feasible space, the minimization of the loss function leads to a degenerate distribution. We summarize the results based on the fitted lognormal distribution.

The fitted lognormal distribution is highly skewed right, with a median survival time of 32 months. This is consistent with the fact that 206 of 504, or 41% of, devices had failed by the end of study date in month 36 (recall that many did not begin service until well after month one) and the declining number of failures in later months. As a check for the credibility of the point estimates and prediction intervals, the procedure of section 2 was used with the lognormal distribution on the first 26 months of observed values (*i.e.*, the data were artificially censored at month 26) in order to obtain point estimates and prediction intervals for months 27–36 for which we have observed values for comparison. Figure 5 and Table 6 show that the point estimates and prediction intervals are quite reasonable: one of ten observed values falls slightly outside the prediction interval, while most observed values are well centered within the prediction interval.

5. Concluding Remarks

The results from our simulation study suggest that the proposed method of estimation performs well when the underlying survival times can be modeled by a lognormal distribution, compared to a Weibull or gamma distribution. There does not seem to be a simple explanation in terms of the tails of the distribution: from Figure 2, all three distributions have comparable tails on the right; on the left, the lognormal has the lightest tail while the Weibull has the heaviest, yet the results corresponding to data generated from the Weibull distribution were better than those for data generated from the gamma distribution. More extensive simulations using a wide variety of

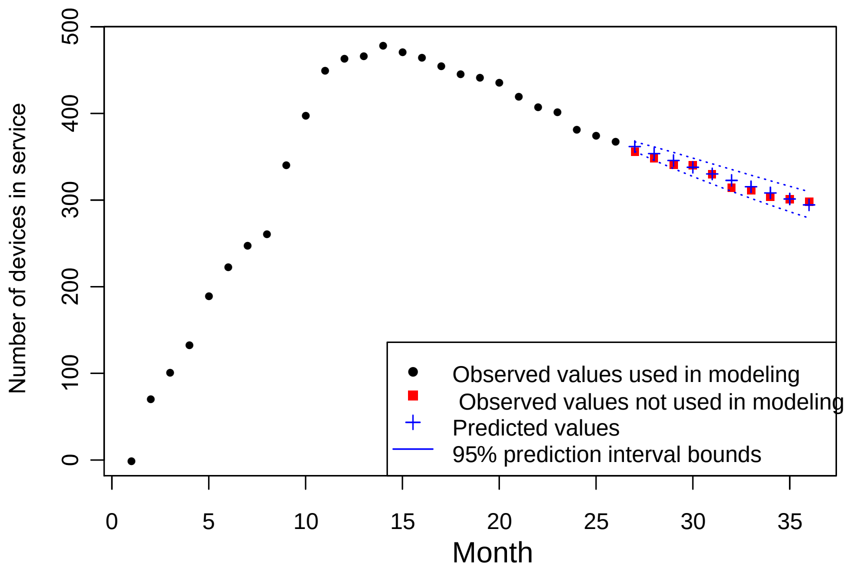

Figure 5. Observed and predicted number of devices in service, along with prediction intervals, by month.

Table 6. Observed and predicted number of units in service in month l_0 under squared error loss and artificial censoring at 26 months, based on a sample size of 504, for the data set of the case study described in section 4.

l_0	$n_l(l_0)$	$\bar{n}_l(l_0)$	95% prediction interval
27	356	361.7	(356.2, 367.3)
28	348	353.6	(345.9, 361.4)
29	341	345.7	(336.3, 355.0)
30	340	337.8	(327.2, 348.5)
31	330	330.2	(318.5, 341.9)
32	314	322.7	(310.1, 335.3)
33	311	315.4	(302.0, 328.7)
34	304	308.2	(294.2, 322.3)
35	301	301.2	(286.6, 315.9)
36	298	294.4	(279.2, 309.6)

distributions should aid in a better understanding of what aspects of the underlying survival distribution affect the performance of the proposed method. With regard to the proposed method of prediction, the results from our simulation study indicate that the use of 'true-model' prediction error variance resulted in a consistent under-coverage; nevertheless, the results from the case study are encouraging. Finally, as with any parametric method, the proposed method relies on a very specific assumption about the underlying distribution from which the data originate. The impact of a misspecified distribution will be a topic for future research.

References

[1] Kaplan, E. I. and Meier, P. 1958. Nonparametric estimation from incomplete observation. *Journal of the American Statistical Association*, 53, 282: 457–481. doi: 10.2307/2281868

[2] Box, G. E. B., Jenkins, G. M., Reinsel, G. C., and Ljung, G. M. 2015. "*Time Series Analysis: Forecasting and Control*", 5th ed., Wiley, New Jersey.

[3] Chatfield, C. 2003. "*The Analysis of Time Series: An Introduction*", 6th ed., Chapman and Hall, London.

[4] Wei, W. S. 2006. "*Time Series Analysis: Univariate and Multivariate Methods*", 2nd ed., Pearson, Boston.

[5] Chatfield, C. 2001. Prediction intervals for time-series forecasting. In "*Principles of Forecasting: A Handbook for Researchers and Practitioners*", J.S. Armstrong, ed., Springer-Verlag, New York.