Developments in Applied Statistics
Anuška Ferligoj and Andrej Mrvar (Editors)
Metodološki zvezki, 19, Ljubljana: FDV, 2003

Estimation of Lifetime Distribution from Incomplete Data and Follow Ups

Ahmed A. Abu-Taleb[1]

Abstract

The motivation for the problem considered in this paper comes from a survival analysis problem in quality-assurance systems that are becoming more popular. For example, when an automobile failures occur within the automotive warranty period, a manufacturer can develop a record of mileage to failure from owners' request for repair. When no failure occur during the warranty period the owner naturally will not report the mileages, and it may be inferred that no record of failures. By using a follow-up survey data can be acquired to include a partial record of nonfailures. A method of estimating life time parameters is proposed for analyzing this kind of data under various scenarios assuming an Exponential lifetime distribution.

1 Introduction

In many statistical estimation problems, the data may not be exact for many reasons. For example, in medical and reliability studies (Collett, 1994; Cox and Oakes, 1984), it is impossible or inconvenient to get complete measurements on all individuals of a random sample. Censored observations are observations that contain partial information about the random variable X of interest. The data that contain censored observations are called incomplete or censored data.

Censoring could be of three types (Lawless, 1982). First, an observation X is said to be right-censored at U if the exact value of the observation is not known, but only that it is greater than or equal to U. Second, an observation X is said to be left-censored at L if it is known only that the observation is less than or equal to L. Third, an observation X is to be interval-censored if it is known only that the observation is between two values, say L and U, where L < U. In some applications, some observations are censored on the right and some on the left. In this case, the data are said to be doubly censored.

[1] Department of Mathematics and Statistics, Jordan University of Science and Technology, P.O. box 3030 Irbid, Jordan; e-mail: abutaleb@just.edu.jo

One of the main goals in Survival Analysis is to estimate the distribution function of the lifetime X of individuals in a population from the observed lifetimes of n individuals. Once the estimator is derived, the focus shifts to examining the properties of the estimator.

2 Description of the problem

The motivation for the problem considered comes from a survival analysis problem in quality-assurance systems that are becoming more popular. For example, some automobile manufacturers offer warranties on every new automobile. The warranty is either for a fixed period of calendar time or for a fixed mileage. To monitor the reliability of the new car, the manufacturer is interested in estimating the distribution of total mileage covered before the first breakdown occurs. The manufacturer has a clear definition of what constitutes a breakdown in a car. Presumably, failures attributable to manufacturing defects are treated as breakdowns. Assume that the warranty is offered for a fixed calendar time. Suppose that a random sample of new cars is being monitored to estimate the mileage distribution. The data generated will have the following structure:

1. If an automobile in the sample breaks down during its warranty period, the manufacturer will know the total number of miles driven on the car when it is brought in for repair.
2. If an automobile has no breakdowns during its warranty period, the manufacturer has no idea how many miles were driven on the car before the first breakdown occurs. It is extremely difficult to get information about the mileage on the car once the warranty period is over. It is relatively easy to get information on miles Z driven on the car at the end of the warranty period.

For an automobile randomly selected, let X denote the number of miles driven before the first breakdown occurs. Let

$$\Delta = \begin{cases} 1 & \text{if the automobile breaks down during the warranty period,} \\ 0 & \text{if the automobile fails to break down during the warranty period,} \end{cases}$$

and

$$Y = \begin{cases} X & \text{if } \Delta = 1, \\ Z & \text{if } \Delta = 0. \end{cases}$$

It is clear that Y = Min{X, Z}, and observing the event {Δ = 0} is equivalent to observing the event {X > Z}. The miles Z driven on the car at the end of the warranty period may not be known. Suzuki (1985a, b) suggested that the manufacturer should pursue a fixed proportion p of owners in the sample and get information on Z if the event {Δ = 0} occurs. Under this particular scenario, Suzuki (1985a, b) derived estimators of the distribution of X both in the parametric and nonparametric frameworks.

In this paper, the following more natural scenario will be considered.

Scenario: Instead of pursuing a fixed proportion p of owners in the sample to get information on Z if the event $\{\Delta = 0\}$ occurs, we contact all the owners in the sample either by telephone or by mailing a questionnaire. The basic information requested is the number Z of miles driven on the car at the conclusion of the warranty period if the car has not broken down during the warranty period. The owner may or may not report the requested information. Let us introduce the random entity,

$$D = \begin{cases} 1 & \text{if the owner reports the requested information,} \\ 0 & \text{if the owner does not report the requested information.} \end{cases}$$

In this scenario, the proportion p of automobiles providing the requested information is random. The random variable D is defined only when the event $\{\Delta = 0\}$ occurs. The data consist of realizations of (Y, Δ, D).

Using the data obtainable in this environment, we will pursue the estimation of the distribution of X in some parametric frameworks.

3 Estimation of lifetime distribution

In this paper, the estimation problem of the survival function of an industrial product in which the real operating time differs from its actual calendar time is considered.

Assume that the warranty is offered for a fixed calendar time. There are two ways to generate data. One could run a sample of automobiles in an industrial laboratory for a fixed length of time commensurate with the warranty period offered. The total number of miles on the odometer is recorded for each automobile if a breakdown occurs. This is how the lab data are generated. Another way is to sell a sample of automobiles to customers and monitor the performance of the automobiles over their warranty periods. This is how the field data are generated. It has been argued that field data are more reliable than the lab data in truly assessing the reliability of automobiles. The problem considered in this paper is how to analyze the field data. Suppose that a random sample of size n of new automobiles is being monitored to estimate the mileage distribution. The data generated will have the following structure:

1. If an automobile in the sample breaks down during its warranty period, the manufacturer will know the total number of miles X driven on the automobile when it is brought in for repair.

2. If an automobile has no breakdowns during its warranty period, the manufacturer has no idea how many miles were driven on the automobile before the first breakdown occurs.

Once the warranty period is over during which no breakdowns occurred, it is difficult to track the owner to obtain the value of X, the number of miles driven on the automobile before the first breakdown occurs. On the other hand, it is feasible to get information on the number Z of miles driven on the automobile at the conclusion of the warranty period. Evidently, X > Z. The focus is now on the manner in which this information is obtained.

Suzuki (1985a, b) proposed the following scheme. Choose and fix an integer $1 \le r \le n$, where n is the sample size. Identify r owners in the sample. No randomization is done in choosing the r owners. Designate this sub-sample as the follow-up sample. If the automobile of an owner in the follow-up sample has no breakdown during the warranty period, record the number Z of miles driven on the automobile at the conclusion of the warranty period. Suzuki assumes that the number Z is always obtainable for every automobile in the follow-up sample.

Let

$$D_i = \begin{cases} 1 & \text{if the i - th owner is designated to be followed up,} \\ 0 & \text{otherwise.} \end{cases}$$

$\frac{r}{n} = \frac{1}{n}\sum_{i=1}^{n} D_i = p^*$. D_i is non-random, and p* is deterministic.

Suzuki (1985a, b) estimated the distribution of X under this scenario in both parametric and nonparametric environments. In Suzuki's scheme, it is possible that all the automobiles in the follow-up sample broke down before the expiration of the warranty period and the remaining (n-r) automobiles had no breakdowns during the warranty period. In such a case, we have no information at all about the remaining (n-r) automobiles. But the probability that such an event will occur is small. In this chapter, we will consider a more natural scenario. If an automobile in the sample has no breakdown during the warranty period, we solicit from the owner information on the number Z of miles driven on the automobile. The owner may or may not provide the solicited information. Let r* be the number of owners who provide the requested information. The entity r* is random. It is in this modified scenario that we derive the maximum likelihood estimator of the distribution of X. Some of the properties of the estimator are also examined.

3.1 Parametric estimation of lifetime distribution

In this section, the estimation of lifetime parameters from incomplete data and follow-ups is considered. In Section 3.1.1, the notation is set up, and assumptions are spelled out. In Section 3.2, the estimation of the lifetime parameters is discussed in general terms. In Section 3.3, the derivation of the estimators is specialized when the underlying distribution is exponential. Some of the properties of the derived estimators are discussed in Section 3.4.

3.1.1 Notation and assumptions

The following notation and assumptions are used in this article.

a. (X_i, Z_i), i = 1, 2, ..., n: independent, identically distributed pairs of random variables, with X_i the mileage on the i-th automobile in the sample before the first breakdown occurs and Z_i the total mileage on the automobile at the end of the warranty period.

b. $f(\cdot)$, $S(\cdot)$: the probability density function of X and its corresponding survival function, respectively.

c. $g(\cdot)$, $G(\cdot)$: the probability density function of Z and its corresponding survival function, respectively.

d. Θ : a vector of unknown parameters taking on values in the parameter space Ω. The parameters involved in the density function $f(\cdot)$ and $g(\cdot)$ are indicated by the symbol θ.

e. The observable, for i = 1, 2, ..., n, are

(i) $$\Delta_i = \begin{cases} 1 & \text{if a breakdown occurs before the expiration} \\ & \text{of the warranty period,} \\ 0 & \text{otherwise.} \end{cases}$$

(ii) If $\Delta_i = 1$, let X_i be the mileage shown on the odometer.

(iii) If $\Delta_i = 0$, let

(iv) $$D_i = \begin{cases} 1 & \text{if the owner provides the information on the} \\ & \text{number } Z_i \text{ of miles driven on the car at the} \\ & \text{conclusion of the warranty period,} \\ 0 & \text{otherwise.} \end{cases}$$

(v) $$Y_i = \begin{cases} X_i & \text{if } \Delta_i = 1, \\ Z_i & \text{if } \Delta_i = 0 \text{ and } D_i = 1, \\ \tau & \text{if } \Delta_i = 0 \text{ and } D_i = 0, \end{cases}$$

where the special symbol τ indicates that no information is available on the automobile, except that $X_i > Z_i$ with both X_i and Z_i unobservable.

One needs to emphasize that the random entity D_i is defined only when the event $\{\Delta_i = 0\}$ occurs. The data can be symbolically written as (Δ_i, D_i, Y_i), i = 1, 2, ..., n. The sample space of the data when n = 3 consists of 27 outcomes. For example:

$(1, -, X_1), (1, -, X_2), (1, -, X_3), 0 < X_1, X_2, X_3 < \infty$.

$(1, -, X_1), (1, -, X_2), (0, 1, Z_3), 0 < X_1, X_2 < \infty$ and $0 < Z_3 < \infty$.

$(1, -, X_1), (1, -, X_2), (0, 0, \tau), 0 < X_1, X_2 < \infty$.

$(1, -, X_1), (0, 0, \tau), (1, -, X_3), 0 < X_1, X_3 < \infty$.

$(0, 1, Z_1), (1, -, X_2), (0, 1, Z_3), 0 < X_2 < \infty$ and $0 < Z_1, Z_3 < \infty$.
$(0, 0, \tau), (0, 1, Z_2), (1, -, X_3), 0 < X_3 < \infty$ and $0 < Z_2 < \infty$.
$(0, 0, \tau), (0, 0, \tau), (0, 0, \tau)$.
The lower case letters (y_i, δ_I, d_i) are used to indicate the realizations of (Δ_i, D_i, Y_i), i =1, 2, ..., n. The generic symbol (Δ, D, Y) is used for the data obtainable on an automobile.

f. $n_u = \sum_{i=1}^{n} \Delta_i$ = the number of automobiles which broke down during the warranty period.

g. $n_c = \sum_{i=1}^{n} (1-\Delta_i) D_i$ = the number of automobiles which did not break down during the warranty period, but for which the mileage is obtained through follow-ups.

h. $n_\ell = \sum_{i=1}^{n} (1-\Delta_i)(1-D_i)$ = the number of automobiles which did not break down during the warranty period and for which no information is available on the number of miles driven at the conclusion of the warranty period.

i. $n = n_u + n_c + n_\ell$ = the total number of automobiles in the sample.

Assumptions:

a. X_i and Z_i , i = 1, 2, ..., n are independent for all i.
b. All breakdowns during the warranty will be reported to the manufacturer.
c. $P(D_i = 1 \mid \Delta_i = 0) = p$, unknown.

3.2 Estimation of the lifetime parameters

To derive an estimator of the underlying parameter vector based on the likelihood principle, the contribution of each and every observation to the likelihood has to be worked out. There are basically three types of observations. All these types along with their likelihoods are enumerated below.

Type 1. The automobile breaks down during the warranty period. Let y be the mileage on the odometer at the time of the breakdowns. Its likelihood = f(y) P(Z > y).

Type 2. The automobile has no breakdowns during the warranty period. The number of miles y driven during the warranty period is available. Its likelihood = p g(y) S(y).

Type 3. The automobile has no breakdowns during the warranty period. The number of miles y driven during the warranty period is not available. Its likelihood

$$(1\text{-}p)\,P(X > Z) = (1-p)\int_0^\infty P(X > z)\,g(z)\,dz = (1-p)\int_0^\infty S(z)\,g(z)\,dz.$$

The likelihood of the data (δ_i, d_i, y_i), i = 1, 2, ..., n can be written as

$$L(\Theta) = \prod_{i=1}^{n} \left[f(y_i).G(y_i)\right]^{\delta_i} \left[p.g(y_i)S(y_i)\right]^{(1-\delta_i)d_i} \left[(1-p)P(X > Z)\right]^{(1-\delta_i)(1-d_i)} =$$

$$= p^{n_c}(1-p)^{n_\ell} \left[P(X > Z)\right]^{n_\ell} \prod_{i=1}^{n} \left[f(y_i).G(y_i)\right]^{\delta_i} \left[g(y_i)S(y_i)\right]^{(1-\delta_i)d_i}.$$

A commonsensical estimate $\hat{p}$ of p is

$$\hat{p} = \frac{n_c}{n_c + n_\ell}.$$

One can see that this is the maximum likelihood estimate of p. Differentiating $\ell(\Theta)$ with respect to p and equating it to zero gives

$$\frac{\partial \ell(\Theta)}{\partial p} = \frac{n_c}{p} + \frac{n_\ell}{1-p} = 0.$$

The unique solution of this equation is given by

$$\hat{p} = \frac{n_c}{n_c + n\ell}.$$

To find the maximum likelihood estimate of other components of Θ, we compute the partial derivatives of $\ell(\Theta)$, equate the derivatives to zero, and solve the resultant equations.

3.3 Derivation of the estimators if the underlying distribution is exponential

Let X be distributed as Exp(θ) and Z as Exp(λ). The parameter vector Θ is identified as Θ = (θ, λ, p). Note that:

P(Failure before the warranty period) = P(X ≤ Z) = θ/(λ+θ);

P(Failure after the warranty period but the owner responds)

= p. P(X > Z) = p.λ/(λ+θ);

P(Failure after the warranty period, but the owner does not respond)

= (1-p).P(X > Z) = (1-p).λ/(λ+θ).

The likelihood function for the n observations is given by

$$L(\Theta) = \prod_{i=1}^{n} \left[f(y_i).G(y_i)\right]^{\delta_i} \left[p.g(y_i)S(y_i)\right]^{(1-\delta_i)d_i} \left[(1-p)P(X > Z)\right]^{(1-\delta_i)(1-d_i)}.$$

After some simplification, the likelihood simplifies as

$$L(\Theta) = \prod_{i=1}^{n} \left[\theta\, e^{-\theta y_i}\, e^{-\lambda y_i}\right]^{\delta_i} \left[p\,\lambda\, e^{-\lambda y_i}\, e^{-\theta y_i}\right]^{(1-\delta_i)d_i} \left[\frac{(1-p)\lambda}{\theta+\lambda}\right]^{n_\ell}$$

$$= \theta^{n_u} p^{n_c} \lambda^{n_c} e^{-(\lambda+\theta)\sum_{i=1}^{n}(\delta_i + (1-\delta_i)d_i)y_i} \left[\frac{(1-p)\lambda}{\theta+\lambda}\right]^{n_\ell}.$$

The corresponding log-likelihood function is

$$\ell(\Theta) = \ell n L(\Theta) =$$

$$= \mathrm{n_u} \ell n(\theta) + n_c\left[\ell n(p) + \ell n(\lambda)\right] - (\lambda+\theta)\sum_{i=1}^{n}(\delta_i + (1-\delta_i)d_i)y_i + n_\ell \ell n\frac{(1-p)\lambda}{\theta+\lambda}.$$

The maximum likelihood estimates of λ and θ are found by differentiating this function with respect to λ and θ and equating the derivatives to zero. The resulting equations are

$$\frac{\partial \ell n\ell(\Theta)}{\partial\theta} = \frac{n_u}{\theta} - \sum_{i=1}^{n}(\delta_i + (1-\delta_i)d_i)y_i - \frac{n_\ell}{\theta+\lambda} = 0 \tag{1}$$

and

$$\frac{\partial\ell(\Theta)}{\partial\lambda} = \frac{n_c}{\lambda} - \sum_{i=1}^{n}(\delta_i + (1-\delta_i)d_i)y_i + n_\ell\frac{\theta}{\lambda(\theta+\lambda)} = 0. \tag{2}$$

From Equations (1) and (2), we have

$$\sum_{i=1}^{n}(\delta_i + (1-\delta_i)d_i)y_i + \frac{n_\ell}{\theta+\lambda} = \frac{n_u}{\theta} = \frac{\mathrm{n_c} + n_\ell}{\lambda}$$

It follows that

$$\frac{\theta}{\lambda} = \frac{n_u}{n_c + n_\ell}.$$

Therefore, the maximum likelihood estimates $\hat{\theta}$ of θ and $\hat{\lambda}$ of λ are given by

$$\hat{\theta} = \frac{n_u(1-\frac{n_\ell}{n})}{\sum_{i=1}^{n}(\delta_i + (1-\delta_i)d_i)y_i},$$

and

$$\hat{\lambda} = \frac{(n_c + n_\ell)(1-\frac{n_\ell}{n})}{\sum_{i=1}^{n}(\delta_i + (1-\delta_i)d_i)y_i}.$$

The estimate $\hat{\theta}$ has an intuitive interpretation. The data come in three different forms: (1) Mileages of automobiles which broke down during the warranty period; (2) Reported miles driven on the automobile at the conclusion of the warranty period when they had no breakdowns during the warranty period; and (3) The total number of miles driven on the automobiles at the conclusion of the warranty period is not available.

The denominator of $\hat{\theta}$ is the sum of all mileage figures available in 1 and 2. The numerator n_u is weighted down by a factor of the proportion of automobiles on which information on the mileage is available. If every owner reports censored values Z_i's when needed, then $n_\ell = 0$ and the estimate $\hat{\theta}$ is the standard estimate.

3.4 Properties of the estimators

The asymptotic variance-covariance matrix of the parameters is the inverse of the information matrix, whose elements are found from the second derivatives of the log-likelihood function. We have

$$\frac{\partial^2 \ell(\Theta)}{\partial \theta^2} = -\frac{n_u}{\theta^2} + \frac{n_\ell}{(\theta+\lambda)^2},$$

$$\frac{\partial^2 \ell(\Theta)}{\partial \lambda^2} = -\frac{n_c}{\lambda^2} - n_\ell \frac{\theta(\theta+2\lambda)}{\lambda^2(\theta+\lambda)^2},$$

$$\frac{\partial^2 \ell(\Theta)}{\partial p^2} = -\frac{n_c}{p^2} - \frac{n_\ell}{(1-p)^2},$$

$$\frac{\partial^2 \ell(\Theta)}{\partial \theta \partial \lambda} = \frac{n_\ell}{(\theta+\lambda)^2},$$

and

$$\frac{\partial^2 \ell(\Theta)}{\partial \theta \partial p} = \frac{\partial^2 \ell(\theta)}{\partial \lambda \partial p} = 0.$$

The information matrix is the matrix of negative expected value of these partial derivatives. To evaluate these expectations, we need to know the distribution of n_u, n_c, and n_ℓ. The distribution of each of the random variables n_u, n_c, and n_ℓ is Binomial (n, $\theta/(\lambda+\theta)$), Binomial (n, $p\lambda/(\lambda+\theta)$), and Binomial (n, $(1-p)\lambda/(\lambda+\theta)$), respectively.

The negative expectations of the second order derivatives are given by

$$E\left(-\frac{\partial^2 \ell(\Theta)}{\partial \theta^2}\right) = \frac{\mathrm{n}((\theta+\lambda)^2 - \lambda\theta(1-p))}{\theta(\theta+\lambda)^3}.$$

$$E\left(-\frac{\partial^2 \ell(\Theta)}{\partial \lambda^2}\right) = \frac{\mathrm{n}((\theta+\lambda)^2 - \lambda^2(1-p))}{\lambda(\theta+\lambda)^3}.$$

$$E\left(-\frac{\partial^2 \ell(\Theta)}{\partial p^2}\right) = \frac{\mathrm{n}\lambda}{p(1-p)(\theta+\lambda)}.$$

$$E\left(-\frac{\partial^2 \ell(\theta)}{\partial \theta \partial \lambda}\right) = -\frac{n(1-p)\lambda}{(\theta+\lambda)^3}.$$

4 Results and conclusions

The main goal of this work was to present some new methods of analyzing lifetime data collected under certain sampling schemes.

The problem came from the automobile industry. The problem was to model the mileage distribution of a new line of automobiles before the first breakdown occurs. A standard procedure is to run a sample of automobiles inside an industrial lab and note the mileage on the odometer when a breakdown occurs for an automobile. To speed data collection, hard conditions are created for the automobiles in the lab for a quicker degradation of the automobiles. This process is called accelerated life testing.

This method of data collection is losing approval. Running an automobile inside a lab and running it on the American roads are two different propositions. The mileage distribution based on the running of automobile on real roads should truly reflect the reality. Another reason why accelerated life testing has lost its approval is that there is a great deal of difficulty in modeling the acceleration of the life of the automobile. Suzuki (1985a, b) came up with a sampling scheme:

1. Let X denotes the mileage on the odometer of a randomly chosen automobile at the time of its breakdown.
2. Sell a sample of n automobiles. Each automobile carries a warranty against breakdowns for a fixed number of years.
3. If an automobile in the sample breakdown before the warranty period is over, the owner brings in the automobile for repair, and the value of X is noted down.
4. If an automobile has no breakdowns during the warranty period, the value of X for this automobile will not be known. Let Z denote the number of miles the automobile is driven at the conclusion of the warranty period. $X > Z$.
5. Choose and fix $0 < p < 1$. Identify a proportion p of the buyers of the automobiles in the sample. If an automobile in this targeted group of automobiles has no breakdowns, note the number Z of miles on the odometer of the automobile at the conclusion of the warranty period.

Let us illustrate the data collection procedure by a simple scenario. Suppose $n=10$ and $p=0.5$. The following is one possible set of data.

Automobile identity		Breakdown during the warranty period	X	Z
Targeted	1	Yes	x_1	-
	2	Yes	x_2	-
	3	No	-	z_3
	4	No	-	z_4
	5	Yes	x_5	-
Non-targeted	6	Yes	x_6	-
	7	No	-	-
	8	No	-	-
	9	Yes	x_9	-
	10	Yes	x_{10}	-

Assuming that X has a parametric distribution, Suzuki (1985b) estimated the parameters of the mileage distribution using a certain modified maximum likelihood method.

In this work, we modified the sampling protocol of Suzuki (1985b). Our scheme has the same first four ingredients of Suzuki's scheme. We modify the fifth leg of Suzuki's scheme. In Suzuki's scheme, there is an implicit assumption that the owner will furnish the value of Z when requested. This is not realistic. We believe that the response of the owner will be random. The modified version is like the one presented below.

Id.	Breakdown during the warranty period	X	Response of the owner	Z
1	Yes	x_1	-	-
2	Yes	x_2	-	-
3	No	-	Yes	z_3
4	No	-	No	-
5	Yes	x_5	-	-
6	Yes	x_6	-	-
7	No	-	Yes	z_7
8	No	-	Yes	z_8
9	Yes	x_9	-	-
10	Yes	x_{10}	-	-

If an automobile has no breakdowns during the warranty period, contact the owner to provide the value of Z. The owner may or may not respond. In our scheme, we do not have a targeted group of automobiles.

Under the assumption that X has an Exponential distribution, we derived the maximum likelihood estimator of the parameter of the distribution as well as its asymptotic variance.

References

[1] Collett, D. (1994): *Modelling Survival Data in Medical Research*. London: Chapman and Hall.

[2] Cox, D.R. and Oakes, D. (1984): *Analysis of Survival Data*. New York: Chapman and Hall.

[3] Lawless, J.F. (1982): *Statistical Models and Methods for Lifetime Data*. New York: Wiley.

[4] Suzuki, K. (1985a): Nonparametric estimation of lifetime distributions from a record of failures and follow-ups. *Journal of the American Statistical Association*, **80,** 68-72.

[5] Suzuki, K. (1985b): Estimation of lifetime parameters from incomplete field data. *Technometrics*, **27**, 263-271.