Chapter 6

Survival Analysis

Survival analysis traditionally focuses on the analysis of time duration until one or more events happen and, more generally, positive-valued random variables. Classical examples are the time to death in biological organisms, the time from diagnosis of a disease until death, the time between administration of a vaccine and development of an infection, the time from the start of treatment of a symptomatic disease and the suppression of symptoms, the time to failure in mechanical systems, the length of stay in a hospital, duration of a strike, the total amount paid by a health insurance, the time to getting a high school diploma. This topic may be called reliability theory or reliability analysis in engineering, duration analysis or duration modeling in economics, and event history analysis in sociology. Survival analysis attempts to answer questions such as: what is the proportion of a population which will survive past a certain time? Of those that survive, at what rate will they die or fail? Can multiple causes of death or failure be taken into account? How do particular circumstances or characteristics increase or decrease the probability of survival?

To answer such questions, it is necessary to define the notion of "lifetime". In the case of biological survival, death is unambiguous, but for mechanical reliability, failure may not be well defined, for there may well be mechanical systems in which failure is partial, a matter of degree, or not otherwise localized in time. Even in biological problems, some events (for example, heart attack or other organ failure) may have the same ambiguity. The theory outlined below assumes well-defined events at specific times; other cases may be better treated by models which explicitly account for ambiguous events.

While we are still dealing with a random variable X, that may be characterized by its cumulative distribution function $F^X(x)$ (also often referred to as the lifetime distribution function), because survival analysis is primarily interested in the time until one or more events, the random variable is assumed to be nonnegative (it is supported on $[0,\infty)$) and it is traditionally characterized by the survival function $G^X(x) := \mathbb{P}(X > x) = 1 - F^X(x)$. That is, the survival function is the probability that the time of death is later than some specified time x. The survival function is also called the survivor function or survivorship function in problems of biological survival, and the reliability function in mechanical survival problems. Usually one assumes $G(0) = 1$, although it could be less than 1 if there is the possibility of immediate death or failure. The survival function must be nonincreasing: $G^X(u) \geq G^X(t)$ if $u \geq t$. This reflects the notion that survival to a later age is only possible if all younger ages are attained. The survival function is usually assumed to approach zero as age increases without bound, i.e., $G(x) \to 0$ as $x \to \infty$, although the limit could be greater than zero if eternal life is possible. For instance, we could apply survival analysis to a mixture of stable and unstable carbon isotopes; unstable isotopes would decay sooner or later, but the stable isotopes would last indefinitely.

Typically, survival data are not fully and/or directly observed, but rather censored and the most commonly encountered form is right censoring. For instance, suppose patients are followed in a study for 12 weeks. A patient who does not experience the event of interest for the duration of the study is said to be right censored. The survival time for this person is considered to be at least as long as the duration of the study. Another example of right censoring is when a person drops out of the study before the end of the study observation time and did not experience the event. This person's survival time is said to be right censored, since we know that the event of interest did not happen while this person was under observation. Censoring is an important issue in survival analysis, representing a particular type of modified data.

Another important modification of survival data is truncation. For instance, suppose that there is an ordinary deductible D in an insurance policy. The latter means that if a loss occurs, then the amount paid by an insurance company is the loss minus deductible. Then a loss less than D may not be reported, and as a result that loss is not observable (it is truncated). Let us stress that truncation is an illusive modification because there is nothing in truncated data that manifests about the truncation, and only our experience in understanding of how data were collected may inform us about truncation. Further, to solve a statistical problem based on truncated observations, we need to know or request corresponding observations of the truncating variable.

It also should be stressed that censoring and/or truncation may preclude us from consistent estimation of the distribution of a random variable over its support, and then a feasible choice of interval of estimation becomes a pivotal part of statistical methodology. In other words, censoring or truncation may yield a destructive modification of data.

As it will be explained shortly, survival analysis of truncated and censored data requires using new statistical approaches. The main and pivotal one is to begin analysis of an underlying distribution not with the help of empirical cumulative distribution function or empirical density estimate but with estimation of a hazard rate which plays a pivotal role in survival analysis. For a continuous lifetime X, its hazard rate (also referred to as the failure rate or the force of mortality) is defined as $h^X(x) := f^X(x)/G^X(x)$. Similarly to the density or the cumulative distribution function, the hazard rate characterizes the random variable. In other words, knowing hazard rate implies knowing the distribution. As we will see shortly, using a hazard rate estimator as the first building block helps us in solving a number of complicated problems of survival analysis including dealing with censored and truncated data.

Estimation of the hazard rate has a number of its own challenges, and this explains why it is hardly ever explored for the case of direct observations. The main one is that the hazard rate is not integrated over its support (the integral is always infinity), and then there is an issue of choosing an appropriate interval of estimation. Nonetheless, estimation of the hazard rate (in place of more traditional characteristics like density or cumulative distribution function) becomes more attractive for truncated and/or censored data. Further, as it was mentioned earlier, truncation and/or censoring may preclude us from estimation of the distribution over its support, and then the problem of choosing a feasible interval of estimation becomes bona fide. Further, the hazard rate approach allows us to avoid using product-limit estimators, like a renowned Kaplan–Meier estimator, that are the more familiar alternative to the hazard rate approach. We are not using a product-limit approach because it is special in its nature, not simple for statistical analysis, and cannot be easily framed into our E-estimation methodology. At the same time, estimation of the hazard rate is based on the sample mean methodology of our E-estimation.

The above-presented comments explain why the first four sections of the chapter are devoted to estimation of the hazard rate for different types of modified data. We begin with a classical case of direct observations considered in Section 6.1. It introduces the notion of the hazard rate, explains that the hazard rate, similarly to the density or the cumulative distribution function, characterizes a random variable, and explains how to construct hazard rate E-estimator. Right censored (RC), left truncated (LT), and left truncated and right censored (LTRC) data are considered in Sections 6.2-6.4, respectively. Sections 6.5-6.7

discuss estimation of the survival function and the density. Sections 6.8 and 6.9 are devoted to nonparametric regression with censored data.

In what follows α_X and β_X denote the lower and upper bounds of the support of a random variable X.

6.1 Hazard Rate Estimation for Direct Observations

Consider a nonnegative continuous random variable X. It can be a lifetime, or the time to an event of interest (which can be the time of failure of a device, or the time of an illness relapse), or an insurance loss, or a commodity price. In all these cases it is of interest to assess the risk associated with X via the so-called *hazard rate* function

$$h^X(x) := \lim_{v \to 0} \frac{\mathbb{P}(x < X < x + v | X > x)}{v} = \frac{f^X(x)}{G^X(x)}, \quad G^X(x) > 0, \ \ x \geq 0, \tag{6.1.1}$$

where we use our traditional notation $f^X(x)$ for the probability density of X and $G^X(x) := \mathbb{P}(X > x) = \int_x^\infty f^X(u)du = 1 - F^X(x)$ for the survival (survivor) function, and $F^X(x)$ is the cumulative distribution function of X. If one thinks about X as a time to an event-of-interest, then $h^X(x)dx$ represents the instantaneous likelihood that the event occurs within the interval $(x, x + dx)$ given that the event has not occurred at time x. The hazard rate quantifies the trajectory of imminent risk, and it may be referred to by other names in different sciences, for instance as the failure rate in reliability theory and the force of mortality in actuarial science and sociology.

Let us consider classical properties and examples of the hazard rate. The hazard rate, similarly to the probability density or the survival function, characterizes the random variable X. Namely, if the hazard rate is known, then the corresponding probability density is

$$f^X(x) = h^X(x) e^{-\int_0^x h^X(v)dv} =: h^X(x) e^{-H^X(x)}, \tag{6.1.2}$$

where $H^X(x) := \int_0^x h^X(v)dv$ is the cumulative hazard function, and the survival function is

$$G^X(x) = e^{-\int_0^x h^X(v)dv} = e^{-H^X(x)}. \tag{6.1.3}$$

The preceding identity follows from integrating both sides of the equality

$$h^X(x) = -[dG^X(x)/dx]/G^X(x), \tag{6.1.4}$$

and then using $G^X(0) = 1$. Relation (6.1.2) follows from (6.1.1) and the verified (6.1.3).

A corollary from (6.1.3) is that for a random variable X supported on $[0, b]$, $b < \infty$ we get $\lim_{x \to b} h^X(x) = \infty$. This property of the hazard rate for a bounded lifetime plays a critical role in its estimation.

An important property of the hazard rate is that if V and U are independent lifetimes, then the hazard rate of their minimum is the sum of the hazard rates, that is, $h^{\min(U,V)}(x) = h^U(x) + h^V(x)$. Indeed, we have

$$G^{\min(U,V)}(x) = \mathbb{P}(\min(U, V) > x) = \mathbb{P}(U > x)\mathbb{P}(V > x) = G^U(x)G^V(x), \tag{6.1.5}$$

and this, together with (6.1.4), yield the assertion. This property allows us to create a wide variety of shapes for hazard rates. Another important property, following from (6.1.3) and $G^X(\infty) = 0$, is that the hazard rate is not integrable on its support, that is the hazard rate must satisfy $\int_0^\infty h^X(x)dx = \infty$. This is the reason why hazard rate estimates are constructed for a finite interval $[a, a+b] \subset [0, \infty)$ with $a = 0$ being the most popular choice. Further, similarly to the probability density, the hazard rate is nonnegative and has the same smoothness as the corresponding density because the survival function is always smoother than the density. The last but not the least remark is about scale-location transformation $Z=(X-a)/b$ of the lifetime X. This transformation allows us to study Z on the standard unit interval $[0,1]$ instead of exploring X over $[a,a+b]$. Then the following formulae become useful,

$$G^Z(z)=G^X(a+bz), \quad f^Z(z)=bf^X(a+bz), \tag{6.1.6}$$

and

$$h^Z(z)=bh^X(a+bz), \quad h^X(x)=b^{-1}h^Z((x-a)/b). \tag{6.1.7}$$

Among examples of hazard rates for variables supported on $[0,\infty)$, the most "famous" is the constant hazard rate of an exponential random variable X with the mean $\mathbb{E}\{X\}=\lambda$. Then the hazard rate is $h^X(x)=\lambda^{-1}I(x\geq 0)$ and the cumulative hazard is $H^X(x)=(x/\lambda)I(x\geq 0)$. Indeed, the density is $f^X(x)=\lambda^{-1}e^{-x/\lambda}I(x\geq 0)$, the survival function is $G^X(x)=e^{-x/\lambda}$, and this yields the constant hazard rate. The converse is also valid and a constant hazard rate implies exponential distribution, the latter is not a major surprise keeping in mind that the hazard rate characterizes a random variable. A constant hazard rate has coined the name *memoryless* for exponential distribution. Another interesting example is the Weibull distribution whose density is $f^X(x;k,\lambda)=(k/\lambda)(x/\lambda)^{k-1}e^{-(x/\lambda)^k}I(x\geq 0)$, where $k>0$ is the shape parameter and $\lambda>0$ is the scale parameter. The mean is $\lambda\Gamma(1+1/k)$ with $\Gamma(z)$ being the Gamma function, the survivor function is $G^X(x;k,\lambda)=e^{-(x/\lambda)^k}I(x\geq 0)$, the hazard rate function is $h^X(x;k,\lambda)=(k/\lambda)(x/\lambda)^{k-1}I(x\geq 0)$, and the cumulative hazard is $H^X(x;k,\lambda)=(x/\lambda)^kI(x\geq 0)$. Note that if $k<1$ then the hazard rate is decreasing (it is often used to model "infant mortality"), if $k>1$ then the hazard rate is increasing (it is often used to model "aging" process), and if $k=1$ then the Weibull distribution becomes exponential (memoryless) with a constant hazard rate.

Now we are in a position to formulate the aim of this section. Based on a sample $X_1,X_2,\ldots,X_n$ of size n from the random variable of interest (lifetime) X, we would like to estimate its hazard rate $h^X(x)$ over an interval $[a,a+b]$, $a\geq 0$, $b>0$. Because hazard rate is the density divided by the survival function, and the survival function is always smoother than the density, the hazard rate can be estimated with the same rate as the corresponding density. Furthermore, a natural approach is to use (6.1.1) and to estimate the hazard rate by a ratio between estimates of the density and survival function. We will check the ratio-estimate shortly in Figure 6.1, and now simply note that the aim is to understand how a hazard rate may be estimated using our E-estimation methodology because for censored and truncated data, direct estimation of the density or survival function becomes a challenging problem.

To construct an E-estimator of the hazard rate, according to Section 2.2 we need to suggest a sample mean or a plug-in sample mean estimator of Fourier coefficients of the hazard rate. Remember that on $[0,1]$ the cosine basis is $\{\varphi_0(x)=1,\varphi_j(x)=\sqrt{2}\cos(\pi jx),j=1,2,\ldots\}$. Similarly, on $[a,a+b]$ the cosine basis is $\{\psi_j(x):=b^{-1/2}\varphi_j((x-a)/b),j=0,1,\ldots\}$. Note that the cosine basis on $[a,a+b]$ "automatically" performs the above-discussed transformation $Z:=(X-a)/b$ of X. As a result, we can either work with the transformed Z and the cosine basis on $[0,1]$ or directly with X and the corresponding cosine basis on $[a,a+b]$. Here, to master our skills in using different bases, we are using the latter approach. Suppose that $G^X(a+b)>0$ and write for jth Fourier coefficient of $h^X(x)$, $x\in[a,a+b]$,

$$\theta_j:=\int_a^{a+b}h^X(x)\psi_j(x)dx=\int_a^{a+b}\frac{f^X(x)}{G^X(x)}\psi_j(x)dx$$
$$=\mathbb{E}\{I(X\in[a,a+b])[G^X(X)]^{-1}\psi_j(X)\}. \tag{6.1.8}$$

Also, to shed light on the effect of rescaling a random variable, note that if $\kappa_j:=\int_0^1 h^Z(z)\varphi_j(z)dz$ is the jth Fourier coefficient of $h^Z(z)$, $z\in[0,1]$, then

$$\theta_j=b^{-1/2}\kappa_j. \tag{6.1.9}$$

Assume for a moment that the survival function $G^X(x)$ is known, then according to (6.1.8) we may estimate θ_j by the sample mean estimator

$$\tilde{\theta}_j := n^{-1} \sum_{l=1}^{n} \frac{\psi_j(X_l) I(X_l \in [a, a+b])}{G^X(X_l)}. \tag{6.1.10}$$

This Fourier estimator is unbiased and its variance is

$$\mathbb{V}(\tilde{\theta}_j) = n^{-1} \mathbb{V}\Big(\frac{I(X_l \in [a, a+b]) \psi_j(X_l)}{G^X(X_l)}\Big). \tag{6.1.11}$$

Further, using (1.3.4) and the assumed $G^X(a+b) > 0$ we may conclude that the corresponding coefficient of difficulty is

$$d(a, a+b) := \lim_{n \to \infty} \lim_{j \to \infty} n \mathbb{V}(\tilde{\theta}_j) = b^{-1} \int_a^{a+b} h^X(x) [G^X(x)]^{-1} dx. \tag{6.1.12}$$

The coefficient of difficulty explicitly shows how the interval of estimation, the hazard rate and the survival function affect estimation of the hazard rate. As we will see shortly, the coefficient of difficulty may point upon a feasible interval of estimation.

Of course $\tilde{\theta}_j$ is an oracle-estimator which is based on an unknown survival function (note that if we know G^X, then we also know the hazard rate h^X). The purpose of introducing an oracle estimator is two-fold. First to create a benchmark to compare with, and second to be an inspiration for its mimicking by a data-driven estimator, which is typically a plug-in oracle estimator. Further, in some cases the mimicking may be so good that asymptotic variances of the estimator and oracle estimator coincide.

Let us suggest a good estimator of the survival function that may be plugged in the denominator of (6.1.10). Because X is a continuous random variable, the survival function can be written as

$$G^X(x) := \mathbb{P}(X > x) = \mathbb{P}(X \geq x) = \mathbb{E}\{I(X \geq x)\}, \tag{6.1.13}$$

and then its sample mean estimator is

$$\hat{G}^X(x) := n^{-1} \sum_{l=1}^{n} I(X_l \geq x). \tag{6.1.14}$$

Note that $\min_{k \in \{1,\ldots,n\}} \hat{G}^X(X_k) = n^{-1} > 0$ and hence we can use the reciprocal of $\hat{G}^X(X_l)$. The sample mean estimator (6.1.14) may be referred to as an empirical survival function.

We may plug (6.1.14) in (6.1.10) and get the following data-driven estimator of Fourier coefficients θ_j of the hazard rate $h^X(x)$, $x \in [a, a+b]$,

$$\hat{\theta}_j := n^{-1} \sum_{l=1}^{n} \frac{\psi_j(X_l) I(X_l \in [a, a+b])}{\hat{G}^X(X_l)}. \tag{6.1.15}$$

This is the proposed Fourier estimator, and it is possible to show that its coefficient of difficulty is identical to (6.1.12). We may conclude that the empirical survival function is a perfect estimator for our purpose to mimic oracle-estimator (6.1.10). Further, the asymptotic theory shows that no other Fourier estimator has a smaller coefficient of difficulty, and hence the proposed Fourier estimator (6.1.15) is efficient. In its turn this result yields asymptotic efficiency of a corresponding hazard rate E-estimator $\hat{h}^X(x)$.

Let us present an example where the coefficient of difficulty is easily calculated. Consider X with exponential distribution and $\mathbb{E}\{X\} = \lambda$. Then $h^X(x) = 1/\lambda$, $G^X(x) = e^{-x/\lambda}$, and hence

$$d(a, a+b) = b^{-1}\int_a^{a+b} \lambda^{-1}e^{x/\lambda}dx = b^{-1}e^{a/\lambda}[e^{b/\lambda} - 1]. \quad (6.1.16)$$

Note that the coefficient of difficulty increases to infinity exponentially in b. This is what makes estimation of the hazard rate and choosing a feasible interval of estimation so challenging. On the other hand, it is not difficult to suggest a plug-in sample mean estimator of the coefficient of difficulty,

$$\hat{d}(a, a+b) := n^{-1}b^{-1}\sum_{l=1}^{n} I(X_l \in [a, a+b])[\hat{G}^X(X_l)]^{-2}. \quad (6.1.17)$$

To realize that this is indeed a plug-in sample mean estimator, note that (6.1.12) can be rewritten as

$$d(a, a+b) = b^{-1}\int_a^{a+b} f^X(x)[G^X(x)]^{-2}dx = b^{-1}\mathbb{E}\{I(X \in [a, a+b])[G^X(X)]^{-2}\}. \quad (6.1.18)$$

This is what was wished to show. The estimator (6.1.17) may be used for choosing a feasible interval of estimation.

Figure 6.1 helps us to understand the problem of estimation of the hazard rate via analysis of a simulated sample of size $n = 400$ from the Bimodal distribution. The caption of Figure 6.1 explains its diagrams. The left-top diagram exhibits reciprocal of the survival function $G^X(x)$ by the solid line and its estimate $1/\hat{G}^X(X_l)$ by crosses. Note that x-coordinates of crosses indicate observed realizations of X, and we may note that while the support is $[0, 1]$, just a few of the observations are larger than 0.85. We also observe a sharply increasing right tail of $1/\hat{G}(x)$ for $x > 0.7$, and this indicates a reasonable upper bound for intervals of estimation of the hazard rate.

Our next step is to look at two density E-estimates exhibited in the right-top diagram. The solid line is the underlying Bimodal density. The dotted line is the E-estimate constructed for interval $[a, a+b] = [0, 0.6]$ and it is based on $N = 211$ observations from this interval. The dashed line is the E-estimate based on all observations and it is constructed for the interval $[0, 1]$ (the support). The subtitle shows ISEs of the two estimates. Because the sample size is relatively large, both estimates are good and have relatively small ISEs. These estimates will be used by the ratio-estimator of the hazard rate.

Diagrams in the second row show us the same estimate $\hat{d}(a, x)$ of the coefficients of difficulty for different intervals. This is done because the estimate has a sharply increasing right tail. In the left diagram we observe the logarithm of $\hat{d}(0, X_l)$, $l = 1, \ldots, n$, while the right diagram allows us to zoom in on the coefficient of difficulty by considering only $X_l \in [a1, a1+b1]$. Similarly to the left-top diagram, we conclude that interval $[0, 0.6]$ may be a good choice for estimation of the hazard rate, and we may try $[0, 0.7]$ as a more challenging one.

Diagrams in the third (from the top) row exhibit performance of the ratio-estimator

$$\check{h}^X(x) = \frac{\hat{f}^X(x)}{\hat{G}^X(x)}, \quad (6.1.19)$$

where $\hat{f}^X$ is an E-estimate of the density. The ratio-estimate is a natural plug-in estimate that should perform well as long as the empirical survival function is not too small. The left diagram shows us the ratio-estimate based on all observations, that is, it is the ratio of the dashed line in the right-top diagram over the estimated survival function shown in the

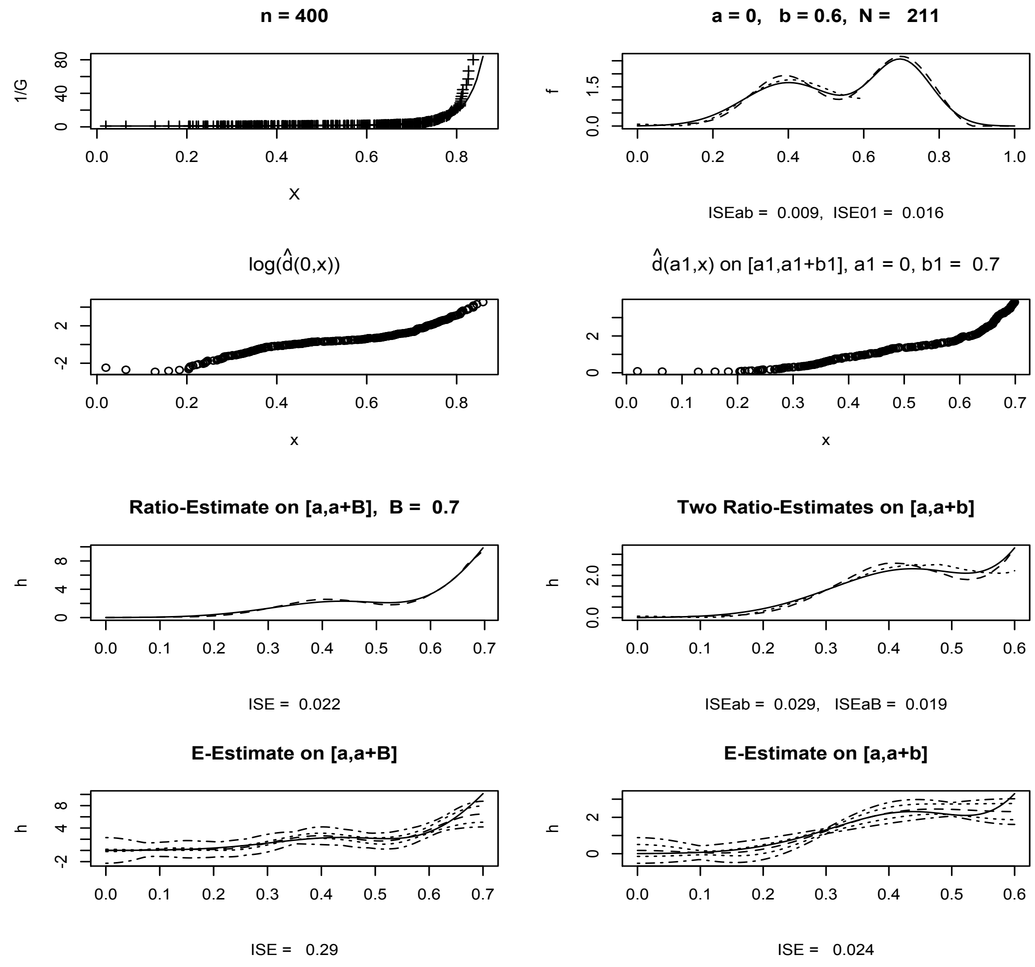


Figure 6.1 *Estimation of the hazard rate based on direct observations. The left-top diagram exhibits reciprocals of the underlying survival function (the solid line) and the empirical one (the crosses at n observed values). The right-top diagram shows the underlying density (the solid line), the E-estimate over $[a, a+b]$ based on observations from this interval (the dotted line), and the E-estimate over the support based on all observations (the dashed line). The second row of diagrams exhibits estimates $\hat{d}(a, X_l)$ over different intervals. In the third row of diagrams, the left diagram shows the hazard rate (the solid line) and its ratio-estimate on $[a, a+B]$, based on all observations, by the dashed line, while the right diagram shows ratio-estimates based on the density estimates shown in the right-top diagram. The left-bottom diagram shows the underlying hazard rate (the solid line) and the E-estimate on $[a, a+B]$ (the dashed line); it also shows the pointwise (the dotted lines) and simultaneous (the dot-dashed lines) confidence bands with confidence level $1-\alpha$, $\alpha = 0.05$. The right-bottom diagram is similar to the left one, only here estimation over interval $[a, a+b]$ is considered. [n = 400, corn = 3, a = 0, b = 0.6, B = 0.7, a1= 0, b1 = 0.7, alpha = 0.05, cJ0 = 4, cJ1 = 0.5, cTH = 4]*

left-top diagram. The estimate is shown only over the interval $[a, a+B] = [0, 0.7]$ because the reciprocal of the estimated survival function is too large beyond this interval. The solid line shows the underlying hazard rate (it increases extremely fast beyond the point 0.7). For this particular simulation the ISE=0.022 is truly impressive. The right diagram shows us two ratio-estimates for the smaller interval $[a, a+b] = [0.0.6]$; the estimates correspond to the two density estimates shown in the top-right diagram. Note that the ratio-estimate based on the density estimate for the interval $[a, a + b]$ (the dotted line) is worse than the ratio-estimate based on the density which uses all observations (the dashed line); also compare the corresponding ISEs in the subtitle. This is clearly due to the boundary effect; on the other hand, the dotted curve better fits the solid line on the inner interval $[0, 0.45]$. It will be explained in Notes at the end of the chapter how to deal with severe boundary effects.

The bottom row shows E-estimates of the underlying hazard rate for different intervals of estimation. The estimates are complemented by pointwise and simultaneous confidence variance-bands introduced in Section 2.6. The E-estimate over the larger interval $[a, a+B] = [0, 0.7]$ is bad. We have predicted the possibility of such outcome based on the analysis of the reciprocal of the survival function and the coefficient of difficulty. Estimation over the smaller interval $[a, a + b] = [0, 0.6]$, shown in the right-bottom diagram, is much better. Further, the E-estimate is better than the corresponding ratio-estimate (the dotted line in the right diagram of the third row) based on the same $N = 211$ observations from this interval. On the other hand, due to the boundary effect, the E-estimate performs worse than the ratio-estimate based on all observations. Further, note that confidence bands present another possibility to choose a feasible interval of estimation.

It is highly advisable to repeat Figure 6.1 with different corner distributions, sample sizes and intervals to get used to the notion of hazard rate and its estimation. Hazard rate is rarely studied in standard probability and statistical courses, but it is a pivotal characteristic in understanding nonparametric E-estimation in survival analysis.

6.2 Censored Data and Hazard Rate Estimation

Censoring is a typical data modification that affects survival data. We begin with the model of right censoring, which more often occurs in applications, and then explain the model of left censoring.

Right censoring (RC) occurs when a case may be removed from a study at some stage of its "natural" lifetime X and the censoring time C of removal is known. It is a standard assumption that both X and C are nonnegative random variables (lifetimes). Under a right censoring model, the smallest among the lifetime of interest X and the censoring time C is observed, and it is also known whether X or C is observed. In other words, instead of a direct sample from X, under the right censoring we observe a sample from a pair of random variables

$$(V, \Delta) := (\min(X, C), I(X \leq C)). \tag{6.2.1}$$

Note that the indicator of censoring Δ tells us when $V = X$ or $V = C$, and because the indicator is always present in data, it is not difficult to realize that we are dealing with data modified by censoring. (The latter is a nice feature of censoring, and in Section 6.3 we will explore another modification, called truncation, when survival data do not manifest the modification.) In what follows we may use the term censoring in place of right censoring whenever no confusion with left censoring occurs.

Let us present several examples that shed light on the right censoring. Suppose that we study a time X from origination of a mortgage loan until the default on the payment; the study is based on data obtained from a large mortgage company. For each originated loan, we get either the time X of the default or the censoring time C when the loan is paid in full due to refinancing, selling a property, final payment, etc. Hence, in this study we observe the default time X only if it is not larger than the censoring time C and otherwise we observe C. Note that we also know when we observe X or C, and hence the example fits model (6.2.1). Another example is the lifetime of a light bulb that may accidentally break at time C before it burns out. But in no way the notion of censoring is restricted to event times. For instance, if an insurance policy has an upper limit C on the payment and there is a claim for a loss with X being the actual loss, then the payment will be the smaller of the loss and the limit. Insurance data typically contain information about payments and policy limits, and hence we are dealing with right censored losses. The latter is a classical example of right-censored data in actuarial science. In a clinical study, typical reasons of right censoring are end of study and withdrawal from study. Note that in all these examples true survival times are equal or greater than observed survival times. As a result, observed data are biased and skewed to the left. Further, estimation of the right tail of distribution may be difficult if not impossible. This is a specific of right censoring that we are dealing with.

Now we are in a position to explain how a hazard rate E-estimator, based on censored data, can be constructed. As usual, the key element is to propose a sample mean estimator for Fourier coefficients. From now on, we are assuming that X and C are independent and continuous random variables. We begin with presenting formulas for the distributions of interest. Because V is the minimum of two independent random variables X and C, there exists a nice formula for its survival function,

$$G^V(v) := \mathbb{P}(\min(X,C) > v) = \mathbb{P}(X > v, C > v) = G^X(v)G^C(v). \tag{6.2.2}$$

Further, for the observed pair (V, Δ) we can write,

$$\mathbb{P}(V \leq v, \Delta = \delta) = \mathbb{P}(\Delta = \delta) - \mathbb{P}(V > v, \Delta = \delta)$$

$$= \mathbb{P}(\Delta = \delta) - \int_v^\infty [f^X(x)G^C(x)]^\delta [f^C(x)G^X(x)]^{1-\delta} dx, \quad v \geq 0,\ \delta \in \{0,1\}. \tag{6.2.3}$$

Differentiation of (6.2.3) with respect to v yields the following formula for the joint mixed density of the observed pair,

$$f^{V,\Delta}(v,\delta) = [f^X(v)G^C(v)]^\delta [f^C(v)G^X(v)]^{1-\delta} I(v \geq 0, \delta \in \{0,1\}). \tag{6.2.4}$$

The formula exhibits a remarkable symmetry with respect to δ which reflects the fact that while C censors X on the right, we may also say that the random variable X also censors C on the right whenever C is the lifetime of interest. In other words, the problem of right censoring is symmetric with respect to the two underlying random variables X and C. This is an important observation because if a data-driven estimator for distribution of X is proposed, it can be also used for estimation of the distribution of C by changing Δ on $1 - \Delta$.

Formula (6.2.4) implies that available observations of X (when $\Delta = 1$) are biased with the biasing function being the survival function $G^C(x)$ of the censored random variable (recall definitions of biased data and a biasing function in Section 3.1). Note that the biasing function is decreasing in v because larger values of X are more likely to be censored. As we know, in general the biasing function should be known for a consistent estimation of an underlying distribution. As we will see shortly, because here we observe a sample from a pair (V, Δ) of random variables, we can estimate the biasing function $G^C(x)$ and hence to estimate the distribution of X. Recall that knowing a distribution means knowing any characteristic of a random variable like its cumulative distribution function, density, hazard rate, survival function, etc. We will see shortly in Section 6.5 that estimation of the density of X is a two-step procedure. The reason for that is that censored data are biased and hence we first estimate the biasing function G^C, which is a complicated problem on its own, and only then may proceed to estimation of the density of X.

Surprisingly, there is no need to estimate the biasing function if the hazard rate is the function of interest (the estimand). This is a pivotal statistical fact to know about survival data where the cumulative distribution function and/or probability density are no longer natural characteristics of a distribution to begin estimation with. The latter is an interesting consequence of data modification caused by censoring.

Let us explain why hazard rate $h^X(x)$ is a natural estimand in survival analysis. Using (6.2.2) and (6.2.4) we may write,

$$h^X(v) := \frac{f^X(v)}{G^X(v)} = \frac{f^{V,\Delta}(v,1)}{G^C(v)G^X(v)} = \frac{f^{V,\Delta}(v,1)}{G^V(v)}. \tag{6.2.5}$$

This is a pivotal formula because (6.2.5) expresses the hazard rate of right censored X via the density and survival function of directly observed variables.

Suppose that we observe a sample of size n from (V, Δ). Denote by $\theta_j := \int_a^{a+b} h^X(x)\psi_j(x)dx$ the jth Fourier coefficient of the hazard rate $h^X(x)$, $x \in [a, a+b]$. Here and in what follows, similarly to Section 6.1, the cosine basis $\{\psi_j(v)\}$ on an interval $[a, a+b]$ is used and recall our discussion of why a hazard rate is estimated over an interval. Assume that $a + b < \beta_V$ (recall that β_V denotes the upper bound of the support of V), and note that using (6.2.5) we can write down a Fourier coefficient as an expectation of a function in V and Δ,

$$\theta_j = \mathbb{E}\Big\{\frac{\Delta I(V \in [a, a+b])\psi_j(V)}{G^V(V)}\Big\}. \tag{6.2.6}$$

This immediately yields the following plug-in sample mean Fourier estimator,

$$\hat{\theta}_j := n^{-1}\sum_{l=1}^{n} \frac{\Delta_l \psi_j(V_l) I(V_l \in [a, a+b])}{\hat{G}^V(V_l)}, \tag{6.2.7}$$

where

$$\hat{G}^V(v) := n^{-1}\sum_{l=1}^{n} I(V_l \geq v) \tag{6.2.8}$$

is the empirical survival function of V.

In its turn, the Fourier estimator implies the corresponding hazard rate E-estimator $\hat{h}^X(x)$. Furthermore, the coefficient of difficulty is

$$d(a, a+b) := \lim_{n\to\infty}\lim_{j\to\infty} n\mathbb{V}(\hat{\theta}_j)$$

$$= b^{-1}\int_a^{a+b} \frac{f^{V,\Delta=1}(v)}{[G^V(v)]^2}dv = b^{-1}\mathbb{E}\Big\{\frac{\Delta I(V \in [a, a+b])}{[G^V(V)]^2}\Big\} \tag{6.2.9}$$

$$= b^{-1}\int_a^{a+b} \frac{h^X(v)}{G^V(v)}dv = b^{-1}\int_a^{a+b} \frac{h^X(v)}{G^X(v)G^C(v)}dv. \tag{6.2.10}$$

Formula (6.2.10) clearly shows how an underlying hazard rate and a censoring variable affect accuracy of estimation. Note that the new here, with respect to formula (6.1.12) for the case of direct observations, is an extra survival function $G^C(v)$ in the denominator. This is a mathematical description of the negative effect of censoring on estimation. Because $G^C(v) \leq 1$ and the survival function decreases in v, that effect may be dramatic. Further, if $\beta_C < \beta_X$ then the censoring implies a destructive modification of data when no consistent estimation of the distribution of X is possible.

Formula (6.2.9) implies that the coefficient of difficulty may be estimated by a plug-in sample mean estimator

$$\hat{d}(a, a+b) := n^{-1}b^{-1}\sum_{l=1}^{n} \Delta_l I(V_l \in [a, a+b])[\hat{G}^V(V_l)]^{-2}. \tag{6.2.11}$$

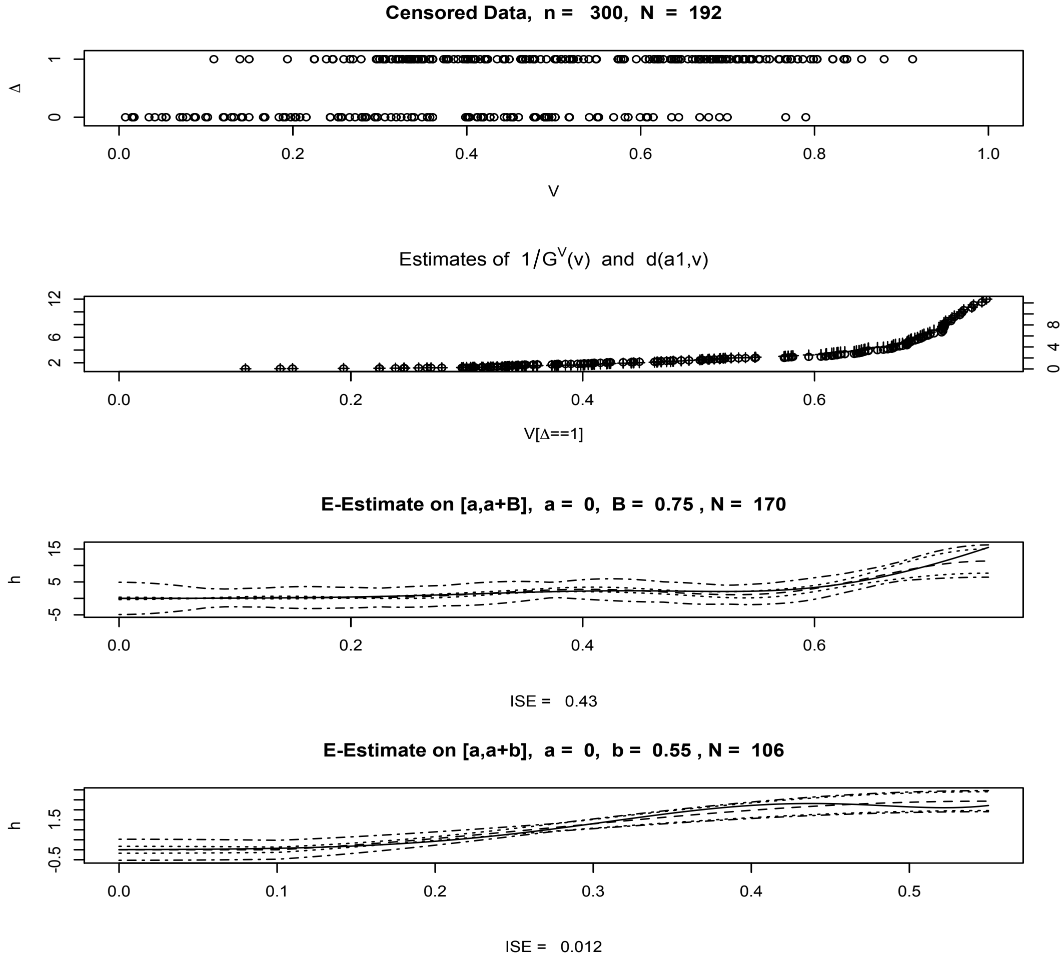


Figure 6.2 *Estimation of the hazard rate based on right censored observations. The top diagram shows* $n = 300$ *observations of* (V, Δ)*, among those* $N := \sum_{l=1}^{n} \Delta_l = 192$ *uncensored ones. Second from the top diagram shows us by crosses values of* $1/\hat{G}(V_l)$ *and by circles values of* $\hat{d}(a1, V_l)$ *for uncensored* $V_l \in [a1, a1 + b1]$*; the corresponding scales are on the left and right vertical axes. Note how fast the right tails increase. The third from the top diagram shows E-estimate of the hazard rate based on observations* $V_l \in [a, a + B]$*, while in the bottom diagram the E-estimate is based on* $V_l \in [a, a + b]$*. A corresponding N shows the number of uncensored observations within an interval. The underlying hazard rate and its E-estimate are shown by the solid and dashed lines, the pointwise and simultaneous* $1 - \alpha$ *confidence bands are shown by dotted and dot-dashed lines.* $\{$*Censoring distribution is either the default Uniform(*$0, u_C$*) with* $u_C = 1.5$*, or Exponential(*λ_C*) with the default* $\lambda_C = 1.5$ *where* λ_C *is the mean. To choose an exponential censoring, set cens = "Expon". Notation for arguments controlling the above-mentioned parameters is evident.*$\}$ *[n = 300, corn = 3, cens = "Unif", uC = 1.5, lambdaC = 1.5, a = 0, b = 0.55, B = 0.75, a1 = 0, b1 = 0.75, alpha = 0.05, cJ0 = 4, cJ1 = 0.5, cTH = 4]*

Note that the new in the estimator, with respect to the case of direct observations of Section 6.1, is that $\hat{G}^X$ is replaced by $\hat{G}^V$. Because $G^V = G^X$ for the case of direct observations, this change is understandable.

The proposed estimator of Fourier coefficients allows us to construct the E-estimator of the hazard rate.

Figure 6.2 sheds light on right censoring, performance of the proposed E-estimator, and the methodology of choosing a feasible interval of estimation explained in Section 6.1. In the particular simulation X is the Bimodal and C is the Uniform(0,1.5). The top diagram shows the sample from $(V, \Delta) = (\min(X, C), I(X \leq C))$. The sample size $n = 300$ and $N := \sum_{l=1}^{n} \Delta_l = 192$ is the number of uncensored realizations of X. The latter indicates a severe censoring. The second from the top diagram allows us to evaluate complexity of the problem at hand and to choose a reasonable interval of estimation. Crosses (and the corresponding scale is shown on the left vertical axis) show values of $1/\hat{G}^V(V_l)$ for $N = 192$ uncensored observations over interval $[a1, a1 + b1] = [0, 0.75]$, while circles show values of the estimated coefficient of difficulty (its scale is on the right-vertical axis). Here we have an interesting outcome that the two functions look similar, and this is due to using different scales and rapidly increasing right tails. The interested reader may decrease $b1$ and get a better visualization of the functions for moderate values of v. Analysis of the two estimates indicates that it is better to avoid estimation of the hazard rate beyond $v = 0.75$, and probably a conservative approach is to consider intervals with the upper bound smaller than 0.6. Let us check this conclusion with the help of corresponding hazard rate E-estimates and confidence bands. The two bottom diagrams allow us to do this, here the chosen intervals are $[a, a + B] = [0, 0.75]$ and $[a, a + b] = [0, 0.55]$, respectively. At first glance, the E-estimate for the larger interval (the third from the top diagram) and its confidence bands may look attractive, but this conclusion is misleading. Indeed, please look at the scale of the bands and realize that the bands are huge! Further, the large bands "hide" the large deviation of the estimate (the dashed line) from the underlying hazard rate (the solid line). Further, the ISE = 0.43, shown in the subtitle, tells us that the estimate is far from the underlying hazard rate. The reason for this is large right tails of $1/\hat{G}^V(v)$ and $\hat{d}(0, v)$ that may be observed in the second from the top diagram. The outcome is much better for the smaller interval of estimation considered in the bottom diagram, and this is despite the smaller number $N = 106$ of available uncensored observations. Also note that the ISE is dramatically smaller.

It is fair to conclude that: (i) The proposed in Section 6.1, for the case of direct observations, methodology of choosing a feasible interval of estimation is robust and can be also recommended for censored data; (ii) The E-estimator performs well even for the case of severe censoring.

Let us finish this section by a remark about left censoring. Under a left censoring, the variable of interest X is censored on the left by a censoring variable C if available observations are from the pair $(V, \Delta) := (\max(X, C), I(X \geq C))$. For instance, when a physician asks a patient about the onset of a particular disease, the answer may be either a specific date or that the onset occurred prior to some specific date. In this case the variable of interest is left censored. Left censoring may be "translated" into a right censoring. To do this, choose a value A that is not less than all available left-censored observations and then consider new observations that are A minus left-censored observations. Then the new observations become right-censored. The latter is the reason why it is sufficient to learn about estimation for right-censored data.

6.3 Truncated Data and Hazard Rate Estimation

It is import to begin this section with a warning. While censoring manifests itself via indicators of censoring and missing manifests itself via empty cells in data, truncation is a "silent" modification of data with no indicators of truncation. Further, truncation is generated by a hidden sequential missing mechanism and we do not even know how many observations were missed. Further, an underlying sequential missing mechanism is MNAR (Missing Not At Random) and it creates biased data. The aim of this section is to explain how to recognize that observations are truncated and then, if possible, to propose an E-estimator of an underlying hazard rate. Left truncation is the most frequently occurred type of truncation, and in what follows we are concentrating on its discussion.

Let us present several examples that shed light on truncation mechanism. Our first example is a classical one in actuarial science. Suppose that we are interested in the distribution of losses incurred by policyholders of an insurance company. Also, suppose that there is no limit on a payment. An insurance company can provide us with data for policies where payments for insurable losses were made. On first glance, this may look like a clear case of direct observations of losses incurred by policyholders, but this is not necessarily the case because a typical insurance policy includes a deductible. The role of a deductible is to decrease the payment (which is the incurred loss minus the deductible) and also to discourage a policyholder from reporting small losses (note that there is no payment on a loss smaller than the deductible, plus there is a possibility of increasing the insurance premium for being a risky policyholder). As a result, smaller (with respect to deductible) incurred losses are missed in the insurance database, and the recorded losses are biased (skewed to the right) with respect to the distribution of underlying losses. Another interesting feature of the MNAR truncation mechanism is that we do not know how many insurable losses occurred. We may say that the underlying incurred loss (the random variable of interest) is left truncated by the deductible (truncating random variable). Let us stress that per se there is nothing in reported losses that manifests their truncated nature. It is up to us to realize that available observations are truncated and that the left truncation causes data to be right-skewed with respect to underlying data because larger realizations of the variable of interest are observed (included in data) with larger likelihood. Further, it is up to us to ask about corresponding observations of the truncating variable. From now on, when the actuarial example is referred to, we are assuming that available data contains reported losses and corresponding deductibles.

Our second example is more challenging. Suppose that we would like to know how long a startup technology company survives until it files for bankruptcy. To study this problem, we can look at all startups that exist at time T, learn about times of their start, and then follow them until their bankruptcy. This gives us data about lifetimes of the startups. Assuming that the process is stationary, it looks like we have desired direct observations of the lifetime of interest. Unfortunately, this is not the case because lifetimes are left truncated by time T at which we search for functioning startup companies. Indeed, if a startup company is bankrupt before time T, it is not included in the study, and furthermore we do not even know that it existed. The example looks more confusing than the actuarial one, but these two examples are similar. Indeed, let us translate the startup example into the actuarial example. We may say that an underlying lifetime of a company is an incurred loss, the time from its start to time T is the deductible, and the observed lifetime of a company is the reported loss. If we use this approach for understanding the startup data, then we may realize that the two examples are similar.

Our third example is from biostatistics where truncation is a typical phenomenon in clinical trials. Suppose that we got data from a study, conducted over the period of last 5 years, about the longevity of patients after a surgery. The data are left truncated because only patients who were alive at the baseline (which is the time of beginning of the study and in our case this was 5 years ago) can participate in the study. Please note how similar this example to the second example with startups, and you may use the above-presented methodology to "translate" it into the actuarial example.

Now, after presenting examples of left truncation, let us describe how truncation modifies underlying data. There are a nonnegative random variable of interest (lifetime) X^* and a truncation random variable T^*. Realizations of the pair (T^*, X^*) are not available directly and are hidden. If the first hidden realization (T_1^*, X_1^*) of (T^*, X^*) satisfies the inequality $X_1^* \geq T_1^*$, that is the random variable of interest is not smaller than the truncation variable, then the pair $(T_1, X_1) := (T_1^*, X_1^*)$ is observed, otherwise this hidden realization is skipped and we even do not know that it occurred. Then the underlying hidden sampling is continued until n realizations of (T, X) are available. Let us look at the truncation one more time via the actuarial example. In the actuarial example X^* is the incurred loss and T^* is the deductible. The incurred loss X^* is reported only if $X^* \geq T^*$ (strictly speaking if $X^* > T^*$ but $\mathbb{P}(X^* = T^*) = 0$ for a continuous X^* and independent T^*), and only then we may observe the reported loss $X = X^*$ and the corresponding deductible $T = T^*$. On the other hand, if $X^* < T^*$, then we do not even know that the loss occurred and also have no information about a corresponding deductible (recall that we have only information about policies with payments, and payments are triggered by reported losses).

The hidden mechanism of collecting data can be described via a negative binomial experiment such that: the experiment stops as soon as nth "success" occurs, data is collected only when a "success" occurs, there is no information on how many "failures" occurred between "successes." Assuming that X^* and T^* are independent, continuous and nonnegative random variables, the probability of "success" can be defined as

$$p := \mathbb{P}(T^* \leq X^*) = \int_0^\infty f^{T^*}(t) G^{X^*}(t) dt. \tag{6.3.1}$$

Here $f^{T^*}(t)$ and $G^{X^*}(t)$ are the density of T^* and the survival function of X^*, respectively. Further, while we do not know the total number N of hidden "failures", the negative binomial distribution sheds some light on it. In particular, the mean and variance of N are calculated as

$$\mathbb{E}\{N\} = n(1-p)p^{-1}, \quad \mathbb{V}(N) = n(1-p)p^{-2}. \tag{6.3.2}$$

Let us make an important remark. If $\mathbb{P}(T^* < c) = 0$ for some positive constant c then all hidden realizations of X^* smaller than c are truncated. As a result, we cannot restore the distribution of X^* for these small values. Hence in general the hazard rate may be estimated only over an interval $[a, a+b]$ with some $a > 0$, and the theory supports this conclusion. Further, as we already know from Section 6.1, in general we cannot estimate the right tail of a hazard rate, and now the left truncation may preclude us from estimating its left tail. This phenomenon will be quantified shortly.

Now we are in a position to present useful probability formulas. For the joint cumulative distribution function of the observed pair (T, X) we may write,

$$F^{T,X}(t, x) := \mathbb{P}(T \leq t, X \leq x) = \mathbb{P}(T^* \leq t, X^* \leq x | X^* \geq T^*) =$$

$$= p^{-1} \mathbb{P}(T^* \leq t, T^* \leq X^* \leq x)$$

$$= p^{-1} \int_0^t f^{T^*}(v) [\int_v^x f^{X^*}(u) du] dv, \quad 0 \leq t \leq x. \tag{6.3.3}$$

Here p is defined in (6.3.1). Then, taking partial derivatives with respect to x and t we get the following expression for the bivariate density,

$$f^{T,X}(t, x) = p^{-1} f^{T^*}(t) f^{X^*}(x) I(0 \leq t \leq x < \infty). \tag{6.3.4}$$

This allows us to obtain, via integration, a formula for the marginal density of X,

$$f^X(x) = f^{X^*}(x)[p^{-1} F^{T^*}(x)]. \tag{6.3.5}$$

In its turn, for values of x such that $F^{T^*}(x) > 0$, (6.3.5) yields a formula for the density of the random variable of interest X^*,

$$f^{X^*}(x) = \frac{f^X(x)}{p^{-1} F^{T^*}(x)} \quad \text{whenever } F^{T^*}(x) > 0. \tag{6.3.6}$$

Note that for values x such that $F^{T^*}(x) = 0$ we cannot restore the density $f^{X^*}(x)$ because all observations of X^* with such values are truncated. In other words, using our notation α_Z for a lower bound of the support of a continuous random variable Z, for consistent estimation of the distribution of X^* we need to assume that

$$\alpha_{X^*} \geq \alpha_{T^*}. \tag{6.3.7}$$

Another useful remark is that formula (6.3.5) mathematically describes the biasing mechanism caused by the left truncation, and according to (3.1.2), the biasing function is $F^{T^*}(x)$. Note that the biasing function is increasing in x.

Let us also introduce a function, which is a probability, that plays a pivotal role in the analysis of truncated data,

$$g(x) := \mathbb{P}(T \leq x \leq X) = \mathbb{P}(T^* \leq x \leq X^*|T^* \leq X^*) = p^{-1}F^{T^*}(x)G^{X^*}(x). \tag{6.3.8}$$

In the last equality we used the assumed independence between T^* and X^*. Note that $g(x)$ is a functional of the distributions of available observations and hence can be estimated. The latter will be used shortly.

Now we have all necessary formulas and notations to explain the method of constructing an E-estimator of the hazard rate of X^*.

Using (6.3.6) and (6.3.8), we conclude that the hazard rate of X^* can be written as

$$h^{X^*}(x) := \frac{f^{X^*}(x)}{G^{X^*}(x)} = \frac{f^X(x)}{\mathbb{P}(T \leq x \leq X)} = \frac{f^X(x)}{g(x)} \text{ whenever } F^{T^*}(x)G^{X^*}(x) > 0. \tag{6.3.9}$$

Note that $h^{X^*}(x)$ is expressed via distributions of available random variables (X, T) and hence can be estimated. Further, note that the restriction $F^{T^*}(x)G^{X^*}(x) > 0$ is equivalent to $g(x) > 0$.

Formula (6.3.9) for the hazard rate is the key for its estimation. Indeed, consider estimation of the hazard rate over an interval $[a, a+b]$ such that $g(x) > 0$ over this interval. Similarly to the previous sections, $\{\psi_j(x)\}$ is the cosine basis on $[a, a+b]$. The proposed sample mean estimator of Fourier coefficients $\theta_j := \int_a^{a+b} \psi_j(x)h^{X^*}(x)dx$ is

$$\hat{\theta}_j := n^{-1}\sum_{l=1}^{n} \frac{\psi_j(X_l)I(X_l \in [a, a+b])}{\hat{g}(X_l)}, \tag{6.3.10}$$

where

$$\hat{g}(x) := n^{-1}\sum_{l=1}^{n} I(T_l \leq x \leq X_l) \tag{6.3.11}$$

is the sample mean estimator of function $g(x)$ defined in (6.3.8). Note that $\hat{g}(X_l) \geq n^{-1}$ and hence this estimator can be used in the denominator of (6.3.10).

The Fourier estimator (6.3.10) yields the corresponding E-estimator of the hazard rate. Further, the corresponding coefficient of difficulty is

$$d(a, a+b) := \lim_{n,j\to\infty} n\mathbb{V}(\hat{\theta}_j)$$

$$= b^{-1}\int_a^{a+b} \frac{h^{X^*}(x)}{g(x)}dx = b^{-1}\mathbb{E}\{I(a \leq X \leq a+b)/g^2(X)\}. \tag{6.3.12}$$

Further, the plug-in sample mean estimator of the coefficient of difficulty is

$$\hat{d}(a, b) := n^{-1}b^{-1}\sum_{l=1}^{n} I(a \leq X_l \leq a+b)/\hat{g}^2(X_l). \tag{6.3.13}$$

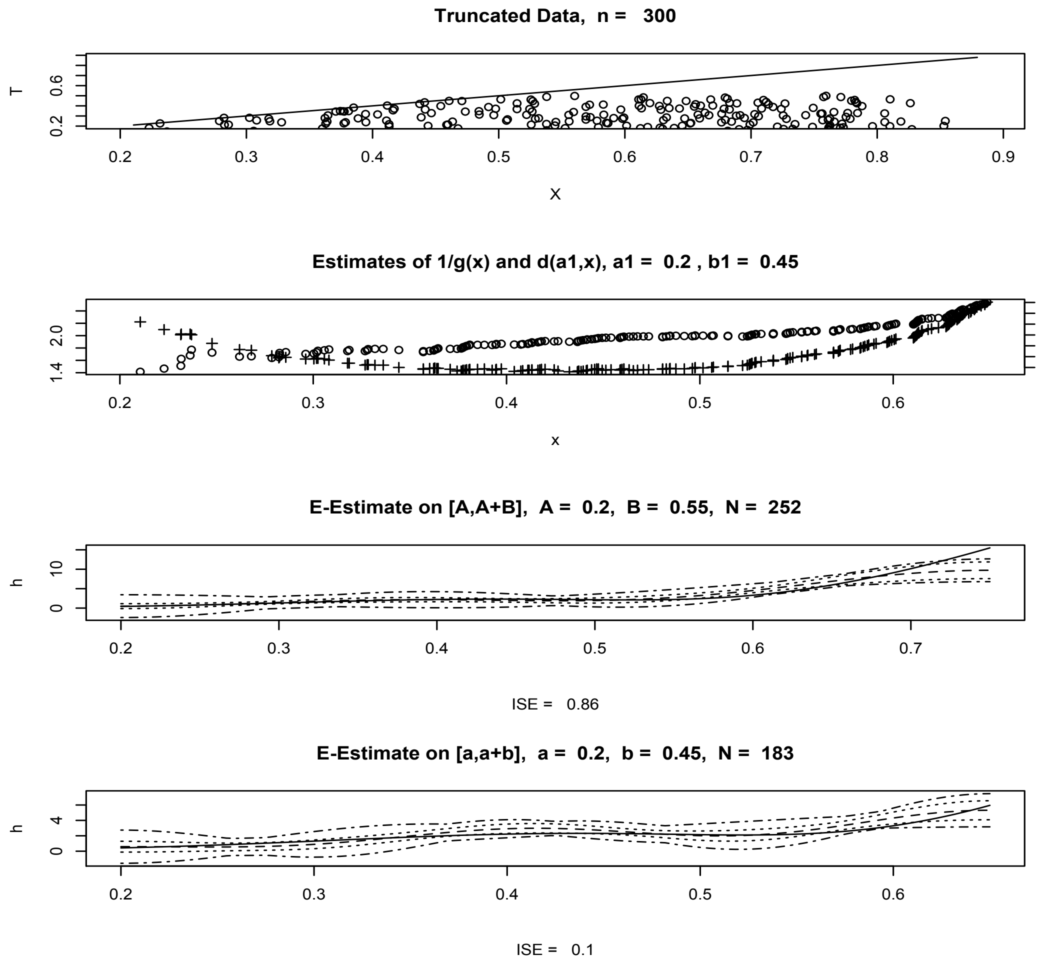


Figure 6.3 *Estimation of the hazard rate based on left truncated observations. The top diagram shows a sample of left truncated observations, the sample size* $n = 300$. *In the simulation* X^* *is the Bimodal and* T^* *is Uniform*$(0, 0.5)$. *Second from the top diagram shows by crosses the estimate* $1/\hat{g}(X_l)$ *and by circles the estimate* $\hat{d}(a1, X_l)$, $X_l \in [a1, a1 + b1]$. *Note the different scales used for these two estimates that are shown correspondingly on the left and right vertical axes. The third from the top diagram shows E-estimate of the hazard rate on interval* $[A, A + B]$, *while in the bottom diagram the E-estimate is for interval* $[a, a + b]$. N *shows the number of observations within a considered interval. The underlying hazard rate and its E-estimate are shown by the solid and dashed lines, the pointwise and simultaneous* $1 - \alpha$ *confidence bands are shown by dotted and dot-dashed lines, respectively.* $\{$*Distribution of* T *is either the default Uniform*$(0, u_T)$ *with* $u_T = 0.5$, *or Exponential(*λ_T*) with the default* $\lambda_T = 0.3$ *where* λ_T *is the mean. Set trunc = "Expon" to choose the exponential truncation. Parameter* α *is controlled by alpha.*$\}$ *[n = 300, corn = 3, trunc = "Unif", uT = 0.5, lambdaT = 0.3, a = 0, b = 0.55, A = 0.2, B = 0.75, a1 = 0.2, b1 = 0.45, alpha = 0.05, cJ0 = 4, cJ1 = 0.5, cTH = 4]*

It is of interest to compare the proposed E-estimator with E-estimators of the previous sections. For the case of direct observations, we have $\mathbb{P}(T^* > 0) = 0$, this yields $g(x) = G^{X^*}(x)$ and hence the E-estimators coincide. For the case of censored observations, the survival function $G^V(x)$ is used in place of $g(x)$, and otherwise the E-estimators are identical. On the other hand, the important new feature of truncated data is that truncation complicates estimation of the left tail of a hazard rate.

Let us look at a particular simulation and analysis of truncated data presented in Figure 6.3. Here X^* is the Bimodal and T^* is the Uniform(0,0.5). The top diagram shows by circles the scattergram of $n = 300$ left-truncated observations of (X, T). Note that all observations are below the solid line $T = X$, this is what we must see in left-truncated observations. Another useful observation is that there are no realizations of X smaller than 0.2. This indicates that the left bound of reasonable intervals of estimation should be larger than 0.2. The second from the top diagram allows us to evaluate complexity of the problem at hand. Crosses (and the corresponding scale is shown on the left vertical axis) show values of the function $1/\hat{g}(X_l)$ for $X_l \in [a1, a1 + b1]$. Note the increasing left and right tails of the function. The coefficient of difficulty $d(a, a + b)$ is now a function of two variables, and the circles show us $\hat{d}(a1, X_l)$, $X_l \in [a1, a1 + b1]$. To see estimates for different intervals, Figure 6.3 allows to change $a1$ and $b1$. Based on the analysis of this diagram, the right side of a feasible interval of estimation should be smaller than 0.6. To test this conclusion, let us choose a relatively large $[A, A + B] = [0.2, 0.75]$ and a smaller $[a, a + b] = [0.2, 0.65]$ interval of estimation. Corresponding hazard rate E-estimates are shown in the two bottom diagrams together with confidence bands. The quality of estimation over the smaller interval is dramatically better, and the confidence bands support our preliminary conclusion about a feasible interval of estimation.

The reader might notice that confidence bands take on negative values while a hazard rate is nonnegative. It is not difficult to make them bona fide (nonnegative), but this will complicate visualization of the left tail of the estimate. Furthermore, these “not bona fide bands” help us to choose and/or justify a feasible interval of estimation.

It is highly recommended to repeat Figure 6.3 with different parameters and underlying variables. Hazard rate is rarely studied in standard probability and statistical classes, and the same can be said about truncation. Analysis of diagrams, generated by Figure 6.3, helps to gain experience in dealing with hazard rate and truncated data.

6.4 LTRC Data and Hazard Rate Estimation

In the previous sections we considered data modified by two mechanisms: right censoring and left truncation. Recall that both these mechanisms imply biased data, with right censoring favoring smaller observations and left truncation favoring larger ones. In many applications these two mechanisms act together. Let us present two examples. We begin with an actuarial example. Suppose that we have data containing payments to policyholders, and the aim is to estimate the distribution of losses incurred by policyholders. If the policies have deductibles and limits on payments, then incurred losses are left truncated by deductibles and right censored by limits on payments. In short, we may say that the available data are left truncated and right censored (LTRC). Now let us consider a clinical trial example where participants, who had a cancer surgery, are divided at the baseline into intervention group (a new medication is given to these participants) and control group (a placebo is given) and then observations of participants continue for a period of five years. The random variable of interest is the lifetime from surgery to death. As we know from the previous sections, the lifetimes in both groups are left truncated by the time from surgery to the baseline, and they may be also right censored by a number of events like the time from surgery to the end of the study, a possible moving to another area, etc. We again are dealing with the LTRC modification.

Let us formally describe a LTRC mechanism of generating a sample of size n of left truncated and right censored (LTRC) observations. There is a hidden sequential sampling from a triplet of nonnegative random variables (T^*, X^*, C^*) whose joint distribution is unknown. T^* is the truncation random variable, X^* is the random variable (lifetime) of interest, and C^* is the censoring random variable. Suppose that (T_k^*, X_k^*, C_k^*) is the kth realization of the hidden triplet and that at this moment there already exists a sample of size $l-1$, satisfying $l-1 \leq \min(k-1, n-1)$, of LTRC observations. If $T_k^* > \min(X_k^*, C_k^*)$ then the kth realization is left truncated meaning that: (i) The triplet (T_k^*, X_k^*, C_k^*) is not observed; (ii) The fact that the kth realization occurred is unknown; (iii) Next realization of the triplet occurs. On the other hand, if $T_k^* \leq \min(X_k^*, C_k^*)$ then the LTRC observation $(T_l, V_l, \Delta_l) := (T_k^*, \min(X_k^*, C_k^*), I(X_k^* \leq C_k^*))$ is added to the LTRC sample whose size becomes equal to l. The hidden sequential sampling from the triplet (T^*, X^*, C^*) stops as soon as $l = n$.

Let us stress two important facts about the model. The former is that the model includes the case when the censoring variable may be smaller than the truncation one, that is it includes the case $\mathbb{P}(C^* < T^*) > 0$. Of course, it also includes the case $\mathbb{P}(C^* \geq T^*) = 1$ with an important example being $C^* := T^* + U^*$ where $\mathbb{P}(U^* \geq 0) = 1$. The latter is that the LTRC mechanism of collecting data can be described via a negative binomial experiment such that the "success" is the event $T^* \leq \min(X^*, C^*)$, the experiment stops as soon as nth "success" occurs, data are collected only when a "success" occurs, there is no information on how many "failures" occurred between "successes", and the probability of "success" is

$$p := \mathbb{P}(T^* \leq \min(X^*, C^*)). \tag{6.4.1}$$

Further, while we do not know the total number of hidden "failures", the negative binomial distribution sheds some light on the hidden number N_f of "failures", and in particular the mean and variance of N_f are calculated as $\mathbb{E}(N_f) = n(1-p)p^{-1}$ and $\mathrm{Var}(N_f) = n(1-p)p^{-2}$.

Now our aim is to understand how the distribution of LTRC observations is related to the distribution of the hidden realizations of the triplet (T^*, X^*, C^*). Set $V^* := \min(X^*, C^*)$. Suppose that a hidden realization of the triplet is observed, meaning that it is given that $T^* \leq V^*$ and we observe (T, V, Δ) where $T = T^*$, $V = V^*$ and $\Delta := \Delta^* := I(X^* \leq C^*)$. Recall that T^* is the underlying truncation random variable, X^* is the lifetime (random variable) of interest, C^* is the underlying censoring random variable, and p, defined in (6.4.1), is the probability of observing the hidden triplet (T^*, V^*, Δ^*). For the joint mixed distribution function of the observed triplet of random variables we can write,

$$\begin{aligned} F^{T,V,\Delta}(t, v, \delta) &:= \mathbb{P}(T \leq t, V \leq v, \Delta \leq \delta) = F^{T^*,V^*,\Delta^*|T^* \leq V^*}(t, v, \delta) \\ &= p^{-1}\mathbb{P}(T^* \leq t, T^* \leq V^* \leq v, \Delta^* \leq \delta), \quad t \geq 0, v \geq 0, \delta \in \{0, 1\}. \end{aligned} \tag{6.4.2}$$

Let us additionally assume that random variables T^*, X^* and C^* are independent and continuous (this assumption will be relaxed later). Then, according to (6.4.1) we get $p = \int_0^\infty f^{T^*}(t)G^{X^*}(t)G^{C^*}(t)dt$, and using (6.4.2) we can write for any $0 \leq t \leq v < \infty$ and $\delta \in \{0, 1\}$,

$$\begin{aligned} &\mathbb{P}(T \leq t, V \leq v, \Delta = \delta) \\ &= p^{-1}\int_0^t f^{T^*}(\tau)\Big[\int_\tau^v f^{X^*}(x)G^{C^*}(x)dx\Big]^\delta \Big[\int_\tau^v f^{C^*}(x)G^{X^*}(x)dx\Big]^{1-\delta} d\tau. \end{aligned} \tag{6.4.3}$$

Taking partial derivatives of both sides in (6.4.3) with respect to t and v yields the following mixed joint probability density,

$$f^{T,V,\Delta}(t, v, \delta) = p^{-1}f^{T^*}(t)I(t \leq v)\Big[f^{X^*}(v)G^{C^*}(v)\Big]^\delta \Big[f^{C^*}(v)G^{X^*}(v)\Big]^{1-\delta}. \tag{6.4.4}$$

Note that the density is "symmetric" with respect to X^* and C^* whenever δ is replaced on $1-\delta$; we have already observed this fact for censored data.

Formula (6.4.4) yields the following marginal joint density

$$f^{V,\Delta}(v, 1) = p^{-1}f^{X^*}(v)G^{C^*}(v)F^{T^*}(v) = h^{X^*}(v)[p^{-1}G^{C^*}(v)F^{T^*}(v)G^{X^*}(v)]. \tag{6.4.5}$$

In the last equality we used definition of the hazard rate $h^{X^*}(x) := f^{X^*}(x)/G^{X^*}(x)$.

The first equality in (6.4.5) yields a nice formula for the density of the random variable of interest,

$$f^{X^*}(x) = \frac{f^{V,\Delta}(x,1)}{p^{-1}G^{C^*}(x)F^{T^*}(x)} \quad \text{whenever } G^{C^*}(x)F^{T^*}(x) > 0. \tag{6.4.6}$$

This formula quantifies the bias in available observations.

Now we are in a position to introduce the probability of an event for available variables which plays a central role in statistical analysis of LTRC data,

$$\begin{aligned} g(x) &:= \mathbb{P}(T \le x \le V) \\ &= \mathbb{P}(T^* \le x \le V^* | T^* \le V^*) \\ &= [p^{-1}G^{C^*}(x)F^{T^*}(x)]G^{X^*}(x), \quad x \in [0, \infty). \end{aligned} \tag{6.4.7}$$

Note that the right side of (6.4.7) contains in the square brackets the denominator of the ratio in (6.4.6). This fact, together with (6.4.6), yields two important formulae. The first one is that the underlying density of X^* may be written as

$$f^{X^*}(x) = \frac{f^{V,\Delta}(x,1)G^{X^*}(x)}{\mathbb{P}(T \le x \le V)} \quad \text{whenever } G^{C^*}(x)F^{T^*}(x) > 0. \tag{6.4.8}$$

Next, if we divide both sides of the last equality by the survival function $G^{X^*}(x)$, then we get the following expression for the hazard rate,

$$h^{X^*}(x) = \frac{f^{V,\Delta}(x,1)}{\mathbb{P}(T \le x \le V)} \quad \text{whenever } G^{C^*}(x)F^{T^*}(x)G^{X^*}(x) > 0, \tag{6.4.9}$$

or equivalently the formula holds whenever $g(x) > 0$ where $g(x)$ is defined in (6.4.7).

The right side of equality (6.4.9) includes characteristics of observed (and not hidden) variables that may be estimated, and this is why we can estimate the hazard rate for values of x satisfying the inequality in (6.4.9). On the other hand, in (6.4.8) the right side of the equality depends on survival function G^{X^*} of an underlying lifetime, and this is why the problem of density estimation for the LTRC is more involved and will be considered later in Section 6.7.

Now we are ready to propose an E-estimator of the hazard rate based on LTRC data. Let $[a, a+b]$ be an interval of estimation where a and b are positive and finite constants. As usual, we begin with a sample mean estimator for Fourier coefficients $\theta_j := \int_a^{a+b} h^{X^*}(x)\psi_j(x)dx$ of the hazard rate $h^{X^*}(x)$, $x \in [a, a+b]$. Recall that $\{\psi_j(x)\}$ is the cosine basis on $[a, a+b]$ introduced in Section 6.1. Using (6.4.9), together with notation (6.4.7), we may propose a plug-in sample mean estimator of the Fourier coefficients,

$$\hat{\theta}_j := n^{-1}\sum_{l=1}^{n} \Delta_l \frac{\psi_j(V_l)}{\hat{g}(V_l)} I(V_l \in [a, a+b]), \tag{6.4.10}$$

where

$$\hat{g}(v) := n^{-1}\sum_{l=1}^{n} I(T_l \le v \le V_l). \tag{6.4.11}$$

Statistic $\hat{g}(v)$ is the sample mean estimator of $g(v)$ (see (6.4.7)) and $\hat{g}(V_l) \ge 1/n$.

Fourier estimator (6.4.10) allows us to construct a hazard rate E-estimator $\hat{h}^{X^*}(x)$, and recall that its construction is based on the assumption that hidden random variables T^*, X^*

and C^* are continuous and mutually independent. The corresponding coefficient of difficulty is

$$d(a,b) := b^{-1}\int_a^{a+b} h^{X^*}(x)g^{-1}(x)dx, \tag{6.4.12}$$

and its plug-in sample mean estimator is

$$\hat{d}(a,b) := n^{-1}b^{-1}\sum_{l=1}^{n}[\hat{g}(V_l)]^{-2}\Delta_l I(V_l \in [a, a+b]). \tag{6.4.13}$$

Figure 6.4 allows us to gain experience in understanding LTRC data, choosing a feasible interval of estimation, and E-estimation. The top diagram shows by circles and triangle the scattergram of LTRC realizations of (T,V). Circles show uncensored observations, corresponding to $\Delta = 1$, and triangles show censored observations corresponding to $\Delta = 0$. Note that all observations are below the solid line $T = V$; this is what we expect from left-truncated observations. The second from the top diagram allows us to evaluate complexity of the problem at hand. Crosses (and the corresponding scale is shown on the left vertical axis) show values of $1/\hat{g}(V_l)$ for uncensored ($\Delta_l = 1$) observations that are used in estimation of Fourier coefficients; see (4.6.10). The estimated coefficient of difficulty is exhibited by circles and its scale is shown on the right vertical axis. Both estimates are shown for $V_l \in [a1, a1 + b1]$, and Figure 6.4 allows one to change the interval. For the data at hand, function $1/\hat{g}(v)$ has sharply increasing tails, and the same can be said about right tail of the coefficient of difficulty. These two estimates can be used for choosing a feasible interval of estimation for the E-estimator of the hazard rate. Two bottom diagrams show us hazard rate E-estimates for different intervals of estimation. Note that the larger interval includes areas with very large values of $1/g(v)$ and this dramatically increases the coefficient of difficulty, the confidence bands, and the ISE. The "effective" number N of uncensored V_l fallen within an interval of estimation is indicated in a corresponding title. In the bottom diagram N is almost twice smaller than in the third from the top diagram, and nonetheless the estimation is dramatically better. This sheds another light on the effect of interval of estimation and the complexity of estimating tails.

So far we have considered the case of continuous and independent underlying hidden random variables. This may be not the case in some applications. For instance, in a clinical trial we may have $C^* := T^* + \min(u, U^*)$ where $\mathbb{P}(U^* \geq 0) = 1$ and u is a positive constant that defines length of the trial. Let us consider a LTRC where a continuous lifetime X^* is independent of (T^*, C^*) while T^* and C^* may be dependent and have a mixed (continuous and discrete) joint distribution.

We begin with formulas for involved distributions. Write,

$$\mathbb{P}(V \leq v, \Delta = 1) = \mathbb{P}(X^* \leq v, X^* \leq C^* | T^* \leq \min(X^*, C^*))$$

$$= p^{-1}\mathbb{P}(X^* \leq v, X^* \leq C^*, T^* \leq X^*) = p^{-1}\int_0^v f^{X^*}(x)\mathbb{P}(T^* \leq x \leq C^*)dx. \tag{6.4.14}$$

Here (compare with identical (6.4.1))

$$p := \mathbb{P}(T^* \leq \min(X^*, C^*)). \tag{6.4.15}$$

Differentiation of (6.4.14) with respect to v yields a formula for the mixed density,

$$f^{V,\Delta}(v,1) = p^{-1}f^{X^*}(v)\mathbb{P}(T^* \leq v \leq C^*) = h^{X^*}(v)[p^{-1}G^{X^*}(v)\mathbb{P}(T^* \leq v \leq C^*)]. \tag{6.4.16}$$

Further, we can write (compare with (6.4.7))

$$\mathbb{P}(T \leq x \leq V) = \mathbb{P}(T^* \leq x \leq V^* | T^* \leq V^*)$$

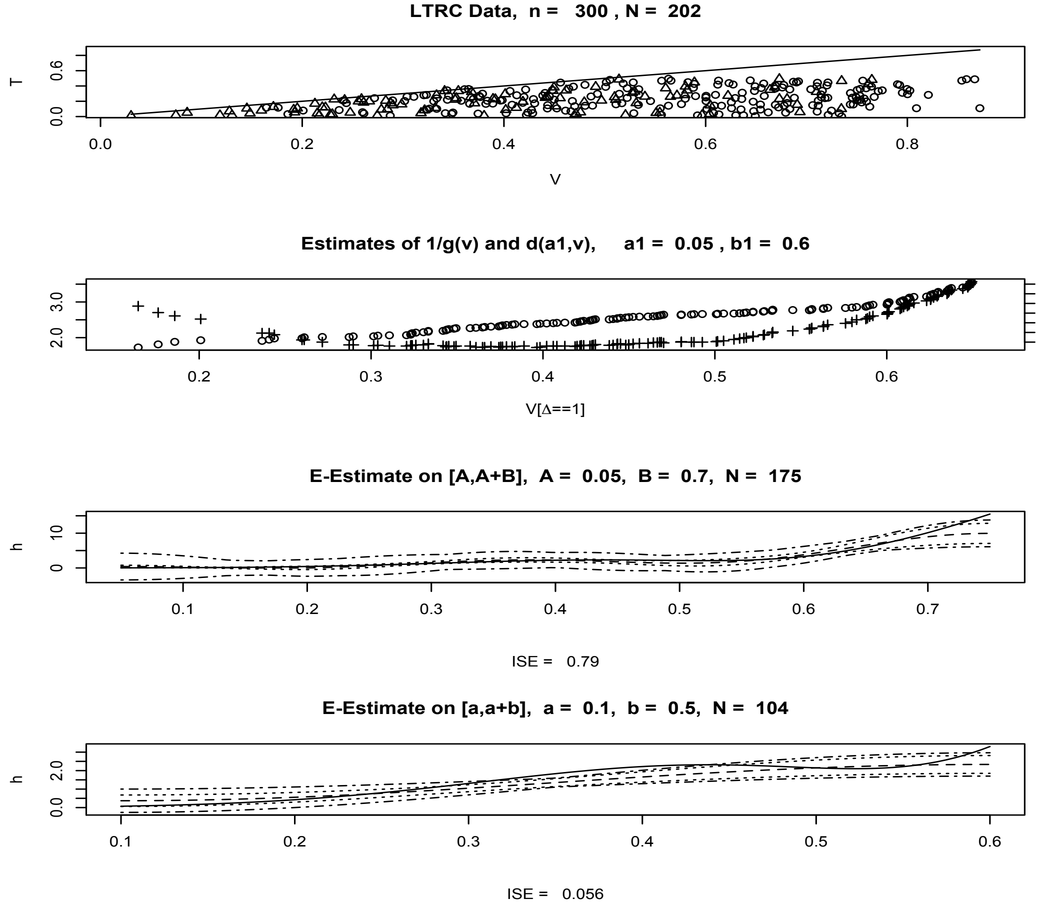


Figure 6.4 *Estimation of the hazard rate based on LTRC observations generated by independent and continuous hidden variables. The top diagram shows a sample of size* $n = 300$ *from* (T, V, Δ)*. Uncensored (their number is* $N := \sum_{l=1}^{n} \Delta_l = 202$*) and censored observations are shown by circles and triangles, respectively. The solid line is* $T = V$*. In the underlying hidden model, the random variable of interest* X^* *is the Bimodal, the truncating variable* T^* *is Uniform*$(0, 0.5)$*, and the censoring variable* C^* *is Uniform*$(0, 1.5)$*. Second from the top diagram shows us by crosses and circles estimates* $1/\hat{g}(V_l)$ *and* $\hat{d}(a1, V_l)$ *for uncensored* $V_l \in [a1, a1 + b1]$*. Note the different scales for the two estimates shown on the left and right vertical axes, respectively. Two bottom diagrams show E-estimate (the dashed line), underlying hazard rate (the solid line) and pointwise and simultaneous* $1 - \alpha = 0.95$ *confidence bands by dotted and dot-dashed lines. The interval of estimation and the number* N *of observations fallen within the interval are shown in the title, the subtitle shows the ISE of E-estimate over the interval. {Distribution of* T^* *is either the default Uniform*$(0, u_T)$ *with* $u_T = 0.5$*, or Exponential(*λ_T*) with the default* λ_T*=0.3 where* λ_T *is the mean. Censoring distribution is either the default Uniform*$(0, u_C)$ *with* $u_C = 1.5$ *or Exponential(*λ_C*) with the default* λ_C *= 1.5. For instance, to choose exponential truncation and censoring, set trunc = "Expon" and cens = "Expon".} [n = 300, corn = 3, trunc = "Unif", uT = 0.5, lambdaT = 0.3, cens = "Unif", uC = 1.5, lambdaC = 1.5, a = 0.1, b = 0.5, A = 0.05, B = 0.7, a1 = 0.05, b1 = 0.6, alpha = 0.05, cJ0 = 4, cJ1 = 0.5, cTH = 4]*

$$= p^{-1}\mathbb{P}(T^* \leq x, V^* \geq x, T^* \leq V^*) = p^{-1}\mathbb{P}(T^* \leq x, C^* \geq x, X^* \geq x)$$

$$= [p^{-1}\mathbb{P}(T^* \leq x \leq C^*)]G^{X^*}(x), \quad x \in [0, \infty). \qquad (6.4.17)$$

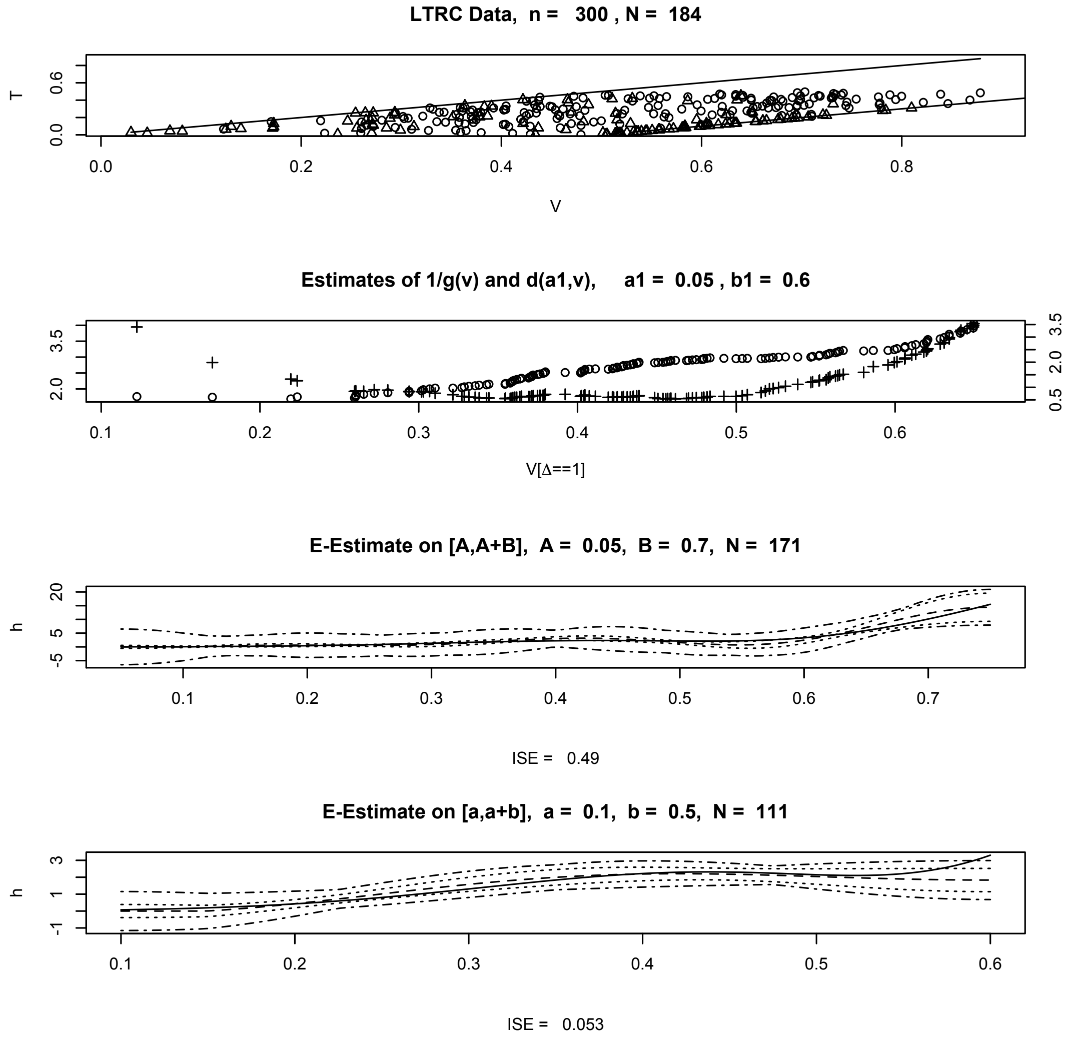


Figure 6.5 *Estimation of the hazard rate for LTRC data with* $C^* = T^* + \min(u, U^*)$ *and* X^* *being independent of* (T^*, C^*). *The default* U^* *is Uniform*$(0, u_C)$ *and independent of* T^*, *and* $u = 0.5$. *The structure of diagrams is identical to Figure 6.4, only in the top diagram the second solid line is added to indicate largest possible censored observations.* {*Distribution of* T^* *is either the default Uniform*$(0, u_T)$ *with* $u_T = 0.5$, *or Exponential*(λ_T) *with the default* λ_T*=0.3 where* λ_T *is the mean. Distribution of* U^* *is either the default Uniform*$(0, u_C)$ *with* $u_C = 1.5$ *or Exponential*(λ_C) *with the default* λ_C *= 1.5. To choose the exponential truncation and censoring, set trunc = "Expon" and cens = "Expon".*} *[n = 300, corn = 3, trunc = "Unif", uT = 0.5, lambdaT = 0.3, cens = "Unif", u = 0.5, uC = 1.5, lambdaC = 1.5, a = 0, b = 0.55, B = 0.75, a1 = 0, b1 = 0.75, alpha = 0.05, cJ0 = 4, cJ1 = 0.5, cTH = 4]*

Combining the last two results we conclude that

$$h^{X^*}(x) = \frac{f^{V,\Delta}(x,1)}{\mathbb{P}(T \le x \le V)} \quad \text{whenever } \mathbb{P}(T^* \le x \le C^*)G^{X^*}(x) > 0. \tag{6.4.18}$$

Note that the inequality in (6.4.18) is equivalent to $\mathbb{P}(T \le x \le V) > 0$. As a result, we may conclude that (6.4.18) is the same formula as (6.4.9) that was established for the case of independent and continuous hidden random variables T^* and C^*. Hence the proposed E-estimator, based on this formula, is robust toward the distribution of (T^*, C^*).

Figure 6.5 allows us to look at LTRC data generated for the case $C^* = T^* + \min(u, U^*)$ and test performance of the hazard rate E-estimator. Please pay attention to the specific shape of the LTRC data in the top diagram. To shed light on the shape, recall the clinical trial example and note that all available observations in a scattergram should be between two parallel lines $T = V$ and $T = V - u$ defined by the baseline and end of the trial, respectively. The second diagram allows us to choose a feasible interval of estimation, and then we have a good chance to obtain a fair hazard rate estimate. Overall, repeated simulations show that the proposed E-estimator performs respectively well and it is robust.

It is highly recommended to repeat Figures 6.4 and 6.5 with different underlying corner distributions and parameters to gain firsthand experience in understanding of and dealing with LTRC data.

6.5 Estimation of Distributions for RC Data

In survival analysis right censoring is the most typical modification of data. For instance, if we are interested in lifetime X of a light bulb that may accidentally break at time C before it burns out, then the lifetime of interest X is right censored by the time of the accident C. Note that RC may be looked at as a missing mechanism when a missed realization of X is substituted by a corresponding realization of C. The missing mechanism is MNAR, because the missing is defined by the value of X, and hence a destructive missing may be expected. As we will see shortly, in some cases RC indeed precludes us from consistent estimation of the distribution of X.

Our aim is to understand when, based on RC observations, a consistent estimation of the distribution of random variables X and C is possible or impossible, and in the former case propose estimators for the survival functions and probability densities.

Statistical model considered in this section is as follows. The available data is a sample from a pair of random variables (V, Δ) where $V := \min(X, C)$ and $\Delta := I(X \leq C)$. It is assumed that X and C are continuous and independent random variables. Introduce a function

$$\begin{aligned} g(v) &:= \mathbb{P}(V \geq v) = G^V(v) = \mathbb{E}\{I(V > v)\} \\ &= \mathbb{P}(V > v) = \mathbb{P}(X > v, C > v) = G^X(v)G^C(v), \end{aligned} \tag{6.5.1}$$

and note that the upper bound of the support of V is

$$\beta_V = \min(\beta_X, \beta_C). \tag{6.5.2}$$

We can make two important conclusions from (6.5.1) and (6.5.2). The first one is that random variables X and C are absolutely symmetric in the sense that if C is the random variable of interest, then it is right censored by X and the corresponding indicator of censoring is $1 - \Delta$. As a result, if we know how to estimate G^X and f^X then the same estimators can be used for the censoring random variable C via using $1 - \Delta$ in place of Δ. Second, the distributions can be estimated only over the interval $[0, \beta_V]$. As a result, if $\beta_C < \beta_X$ then no consistent estimation of the distribution of X is possible, and if $\beta_C > \beta_X$ then no consistent estimation of the distribution of C is possible. This is why RC may be a destructive modification.

Before proceeding to exploring E-estimation for RC data, let us present a canonical in survival analysis Kaplan–Meier estimator of the survival function,

$$\begin{aligned} \tilde{G}^X(x) &:= 1, \; x < V_{(1)}; \quad \tilde{G}^X(x) := 0, \; x > V_{(n)}; \\ \tilde{G}^X(x) &:= \prod_{i=1}^{l-1} [(n-i)/(n-i+1)]^{\Delta_{(i)}}, \quad V_{(l-1)} < x \leq V_{(l)}, \end{aligned} \tag{6.5.3}$$

where $(V_{(l)}, \Delta_{(l)})$ are ordered V_l's with their corresponding Δ_l, $l = 1, \ldots, n$. Kaplan–Meier estimator for G^C is obtained by replacing Δ on $1 - \Delta$.

This is not a simple task to explain the underlying idea of Kaplan–Meier estimator, and it is even more difficult to infer about its statistical properties. Let us present one possible explanation via a specific example. Consider a medical study on longevity of n individuals. Denote by $\tilde{y}_{l_1} < \tilde{y}_{l_2} < \ldots$ observations of V_{l_s} such that corresponding $\Delta_{l_s} = 1$. In other words, $\tilde{y}_{l_s}$ is the sth time of death during the study, and note that because V is a continuous random variable, the probability of simultaneous deaths is zero. To shed additional light on the notation, between times V_{l_1} and V_{l_2}, $l_2 - l_1 + 1$ individuals left the study (their survival times were censored), and after time V_{l_s} only $n - l_s$ individuals are left in the study. Suppose that we are interested in the survival function $G^X(x)$ at a time $x \in [V_{l_2}, V_{l_3})$. Using a probability formula $\mathbb{P}(A \cap B \cap C) = \mathbb{P}(A)\mathbb{P}(B|A)\mathbb{P}(C|A \cap B)$, which expresses the probability of intersection of events via conditional probabilities, we may write,

$$G^X(x) := \mathbb{P}(X > x) = \mathbb{P}\big(\text{survive in } [0, V_{l_1})\big)\mathbb{P}\big(\text{survive in } [V_{l_1}, V_{l_2})|\text{survive in } [0, V_{l_1})\big)$$

$$\times\mathbb{P}\big(\text{survive in } [V_{l_2}, x]|\text{survive in } [0, V_{l_2})\big). \qquad (6.5.4)$$

For the first probability in the right side of (6.5.4) a natural estimate is 1 because no deaths have been recorded prior to the moment V_{l_1}. For the second probability a natural estimate is $(n - l_1)/(n - l_1 + 1)$ because $n - l_1 + 1$ is the number of individuals remaining in the study before time V_{l_1} and then one individual died prior to moment V_{l_2}. Absolutely similarly, for the third probability a natural estimate is $(n - l_2)/(n - l_2 + 1)$. If we plug these estimators in (6.5.4), then we get the Kaplan–Meier estimator

$$\tilde{G}^X(x) := [(n - l_1)/(n - l_1 + 1)] \times [(n - l_2)/(n - l_2 + 1)]. \qquad (6.5.5)$$

This explanation also sheds light on the notion of product-limit estimation often applied to Kaplan–Meier estimator.

Kaplan–Meir estimator is the most popular estimator of survival function for RC data. At the same time, it is not as simple as the classical empirical (sample mean) cumulative distribution estimator $\hat{F}^X(x) := n^{-1}\sum_{l=1}^{n} I(X_l \leq x)$ used for the case of direct observations. Can a sample mean method be used for estimation of the survival function and RC data? The answer is “yes” and below it is explained how this estimator can be constructed.

Let us explain the proposed method of estimation of the survival function and density of X based on RC data. As usual, we begin with formulas for the joint density of observed variables,

$$f^{V,\Delta}(x, 1) = f^X(x)G^C(x) = [f^X(x)/G^X(x)]g(x) = h^X(x)g(x), \qquad (6.5.6)$$

and

$$f^{V,\Delta}(x, 0) = f^C(x)G^X(x) = [f^C(x)/G^C(x)]g(x) = h^C(x)g(x), \qquad (6.5.7)$$

where $g(x)$ is defined in (6.5.1), $h^X(x)$ and $h^C(x)$ are the hazard rates of X and C. Note that (6.5.6) and (6.5.7) contain several useful expressions for the joint density that shed extra light on RC. Further, please look again at the formulas and note the symmetry of the RC with respect to the lifetime of interest and censoring variable.

We begin with the explanation of how the sample mean methodology can be used for estimation of the survival function $G^C(x)$ of the censoring random variable. Recall that it can be estimated only for $x \in [0, \beta]$ where $\beta \leq \beta_V := \min(\beta_X, \beta_C)$. The idea of estimation is based on a formula which expresses the survival function via the corresponding cumulative hazard $H^C(x)$,

$$G^C(x) = \exp\{-H^C(x)\}. \qquad (6.5.8)$$

Using (6.5.1) and (6.5.7), the cumulative hazard may be written as

$$H^C(x) := \int_0^x h^C(u)du = \int_0^x [f^C(u)/G^C(u)]du$$

$$= \int_0^x [f^{V,\Delta}(u,0)/g(u)]du = \mathbb{E}\{(1-\Delta)I(V \in [0,x])/g(V)\}. \quad (6.5.9)$$

As a result, we can use a plug-in sample mean estimator for the cumulative hazard, then plug it in (6.5.8) and get the following estimator of the survival function,

$$\hat{G}^C(x) := \exp\{-n^{-1}\sum_{l=1}^n (1-\Delta_l)I(V_l \le x)/\hat{g}(V_l)\}. \quad (6.5.10)$$

Here

$$\hat{g}(x) := n^{-1}\sum_{l=1}^n I(V_l \ge x) \quad (6.5.11)$$

is the sample mean estimate of $g(x)$ defined in (6.5.1). Note that $\hat{g}(V_l) \ge n^{-1}$ and hence the estimator (6.5.10) is well defined.

The appealing feature of estimator (6.5.10) is its simple interpretation because we estimate the logarithm of the survival function by a sample mean estimator. This is why we may refer to the estimator as a sample mean estimator (it is also explained in the Notes that this estimator, written as a product-limit, becomes a Nelson–Aalen–Breslow estimator which is another canonical estimator in the survival analysis). Another important remark is that $G^X(x)$ also may be estimated by (6.5.10) with $1-\Delta_l$ being replaced by Δ_l.

Now let us consider estimation of the probability density $f^X(x)$, $x \in [0,\beta]$. We are using notation $\{\psi_j(x)\}$ for the cosine basis on $[0,\beta]$. Fourier coefficients of $f^X(x)$ can be written, using (6.5.6), as

$$\theta_j := \int_0^\beta f^X(x)\psi_j(x)dx = \int_0^\beta [f^{V,\Delta}(x,1)\psi_j(x)/G^C(x)]dx$$

$$= \mathbb{E}\{\Delta I(V \le \beta)\psi_j(V)/G^C(V)\}. \quad (6.5.12)$$

This implies a plug-in sample mean estimator

$$\hat{\theta}_j := n^{-1}\sum_{l=1}^n \Delta_l I(V_l \le \beta)\psi_j(V_l)/\hat{G}^C(V_l). \quad (6.5.13)$$

In its turn, the Fourier estimator yields a corresponding density E-estimator $\hat{f}^X(x)$, $x \in [0,\beta]$. Further, if $\beta < \beta_C$ then the corresponding coefficient of difficulty is

$$d(0,\beta) := \beta^{-1}\mathbb{E}\{I(V \le \beta)\Delta[G^C(V)]^{-2}\} = \beta^{-1}\int_0^\beta \frac{f^X(x)}{G^C(x)}dx, \quad (6.5.14)$$

and it can be estimated by a sample mean estimator

$$\hat{d}(0,\beta) := n^{-1}\beta^{-1}\sum_{l=1}^n \Delta_l I(V_l \le \beta_V)[\hat{G}^C(V_l)]^{-2}. \quad (6.5.15)$$

Let us check how the estimator performs. Figure 6.6 exhibits a particular RC sample and the suggested estimates. In the simulation $\beta_C = 1.5 > \beta_X = 1$, and hence the distribution of X can be consistently estimated. Keeping this remark in mind, let us look at the data

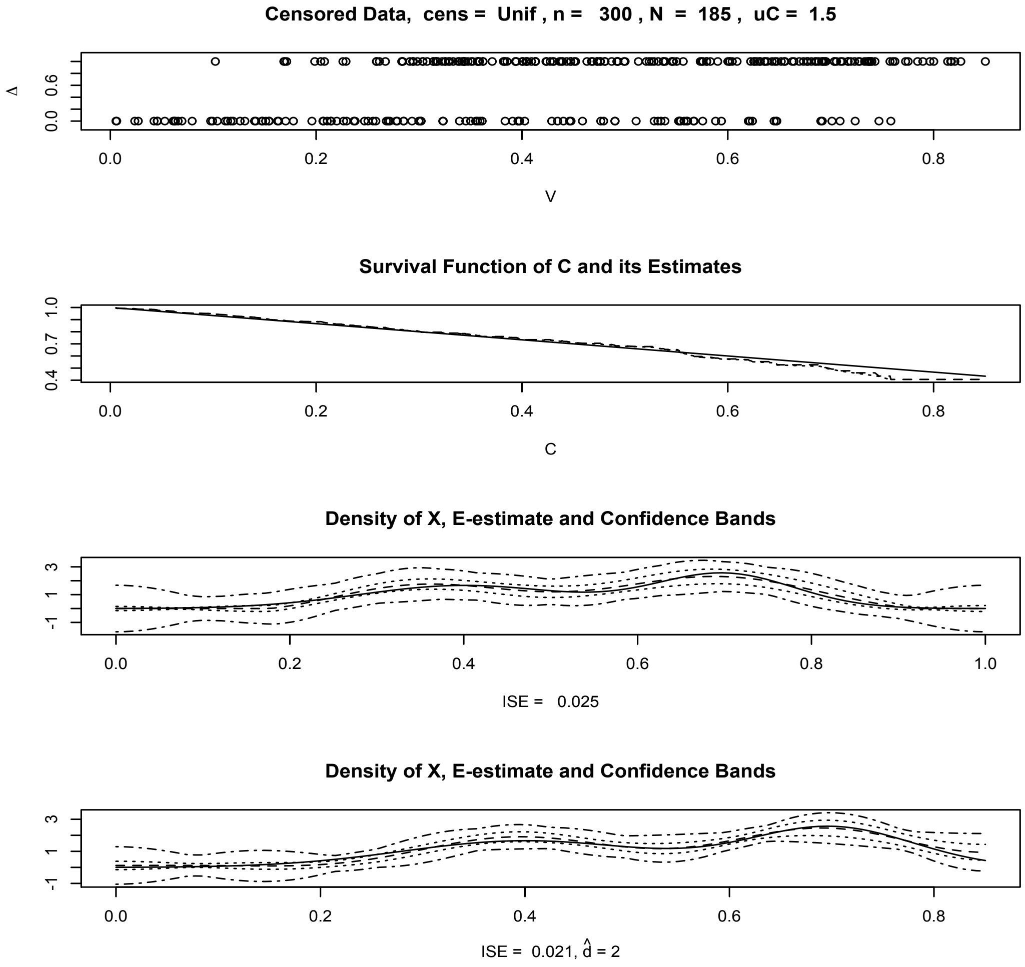
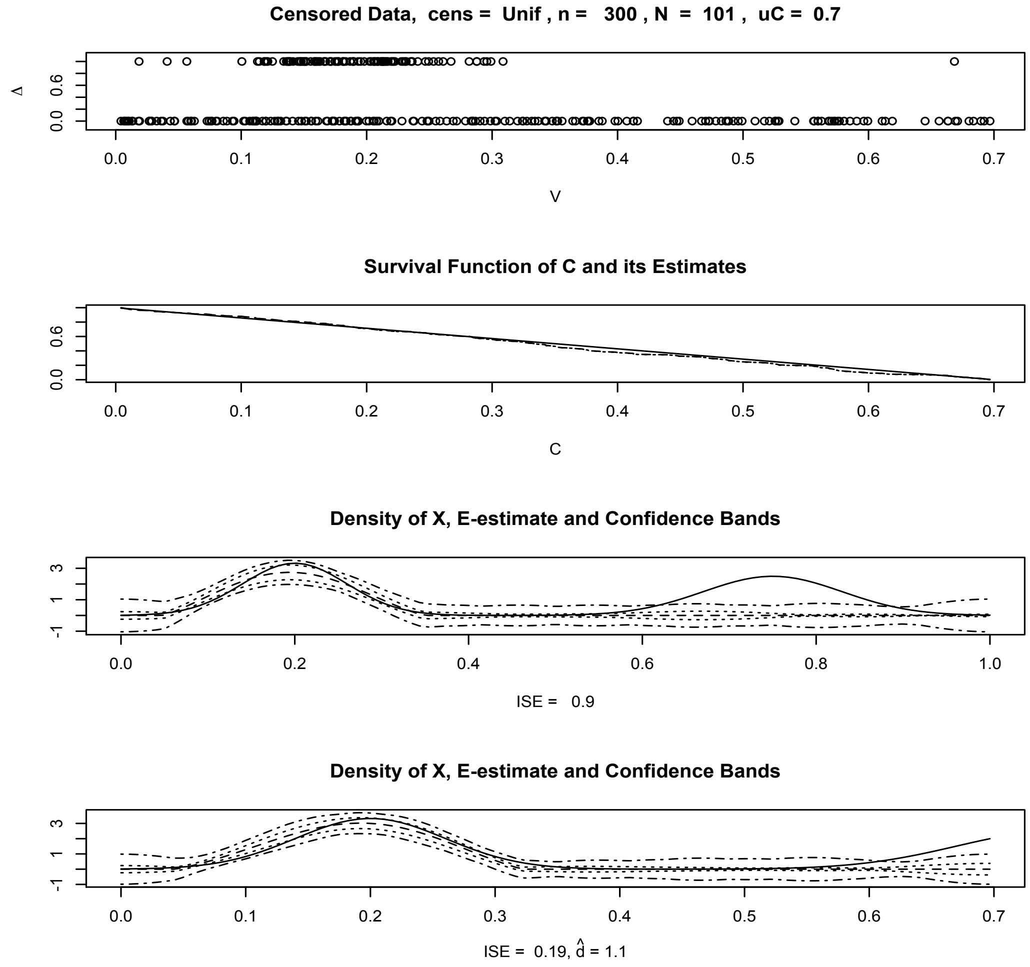


Figure 6.6 *Right-censored data with $\beta_C > \beta_X$ and estimation of distributions. The top diagram shows simulated data. The distribution of the censoring random variable C is uniform on $[0, u_C]$ and the lifetime of interest X is the Bimodal. The second from the top diagram shows by the solid line the underlying G^C, dashed and dotted lines show the sample mean and Kaplan–Meier estimates, respectively. The third from the top diagram shows the underlying density (the solid line), the E-estimate (the dashed line), and $(1-\alpha)$ pointwise (the dotted lines) and simultaneous (the dot-dashed lines) confidence bands. The estimate is for the interval [0,1]. The bottom diagram is similar, only here the estimate is for the interval $[0, V_{(n)}]$ where $V_{(n)} := \max(V_1, \dots, V_n)$. $\{$For exponential censoring set cens = "Expon", and the mean of the exponential censoring variable is controlled by argument lambdaC$\}$. [n = 300, corn = 3, cens = "Unif", uC = 1.5, lambdaC = 1.5, alpha = 0.05, cJ0 = 4, cJ1 = 0.5, cTH = 4]*

and estimates. From the title for the top diagram we note that despite the relatively large sample size $n = 300$, only $N := \sum_{l=1}^{n} \Delta_l = 185$ observations of X are not censored; similarly only $n - N = 115$ observations of C are available. Note that observations of X are shown by the horizontal coordinates of circles corresponding to $\Delta = 1$ and observations of C are similarly shown via $\Delta = 0$. Interestingly, the largest observations of X and C are far from $\beta_V = 1$ despite the large sample size. Further, despite the fact that $\beta_C > \beta_X$, the largest observed lifetime is larger than the largest observed censoring variable. The second from



Censored Data, cens = Unif , n = 300 , N = 101 , uC = 0.7

Survival Function of C and its Estimates

Density of X, E-estimate and Confidence Bands

Density of X, E-estimate and Confidence Bands

Figure 6.7 *Right-censored data with $\beta_C < \beta_X$ and estimation of distributions. This figure is created by Figure 6.6 using arguments uC* $= 0.7$ *and corn* $= 4$. *[n = 300, corn = 4, cens = "Unif", uC = 0.7, lambdaC = 1.5, alpha = 0.05, cJ0 = 4, cJ1 = 0.5, cTH = 4]*

the top diagram shows the underlying G^C, the proposed sample-mean and Kaplan–Meier estimates (consult the caption about the corresponding curves). The estimates are close to each other. Let us also stress that there is no chance to consistently estimate the right tail of the distribution of C due to the fact that $\beta_X < \beta_C$. The two bottom diagrams show how the density E-estimator, constructed for intervals $[0, 1]$ and $[0, V_{(n)}]$, performs. The first interval is of a special interest because $[0, 1]$ is the true support of the density f^X, the second one is a data-driven interval of estimation. Keeping in mind that only $N = 185$ uncensored observations are available and the Bimodal is a difficult density for estimation even for the case of direct observations, the particular density E-estimates are very good. Also look at and compare the corresponding confidence bands and ISEs.

What will be if $\beta_C < \beta_X$? Theoretically this precludes us from consistent estimation of the right tail of $f^X(x)$, but still can we estimate the density over an interval $[0, \beta]$ with $\beta < \beta_V$? The above-presented theory indicates that this is possible, and Figure 6.7 illustrates the case. This figure is generated by Figure 6.6 using $u_C = 0.7$, also note that the underlying density is the Strata. The title for the top diagram shows that only $N = 101$ uncensored observations of the lifetime X are available, and hence we are dealing with a severe censoring. Further, note that while the largest observation of C is close to $\beta_V = \beta_C = 0.7$, all but one observations of X are smaller than 0.32. The second from the top diagram exhibits very good estimates of the survival function $G^C(x)$ over the interval $[0, 0.7]$, and this is not a surprise due to the large number $n - N = 199$ of observed realizations of the censored variable. Now let us look at the two bottom diagrams. As it could be predicted, the density estimate in the second from the bottom diagram is not satisfactory because there are simply no observations to estimate the right part of the density. The E-estimate for the interval $[0, V_{(n)}]$, shown in the bottom diagram, is better but still its right tail is poor because we have just one observation of X which is larger than 0.32 and hence there is no way an estimator may indicate the large underlying right stratum.

It is recommended to repeat Figures 6.6 and 6.7 with different corner functions, censoring distributions and parameters to gain an experience in dealing with RC data.

6.6 Estimation of Distributions for LT Data

The aim is to understand how survival functions and probability densities for random variables, defining a left truncation (LT), may be estimated. Let us briefly recall the LT mechanism and involved random variables (lifetimes), more can be found in Section 6.3. A hidden random variable of interest X^* is observed only if X^* is not smaller than a hidden truncation random variable T^*, otherwise it is not even known that the event $X^* < T^*$ occurred. In what follows it is assumed that X^* and T^* are independent, continuous and nonnegative random variables (lifetimes). Recall notation α_Z and β_Z for the lower and upper bounds of the support of a random variable Z. It is assumed that supports of the hidden variables satisfy the following two relations,

$$\alpha_{X^*} \geq \alpha_{T^*}, \tag{6.6.1}$$

and

$$\beta_{X^*} \geq \beta_{T^*}. \tag{6.6.2}$$

Let us comment on these two assumptions. If $\alpha_{X^*} < \alpha_{T^*}$ then all observations of X^* less than α_{T^*} are truncated and not available. This precludes us from estimation of $f^{X^*}(x)$ for $x < \alpha_{T^*}$. If $\beta_{X^*} < \beta_{T^*}$ then the right tail of $f^{T^*}(t)$ cannot be consistently estimated. Let us also note that in what follows we are interested only in the case $\beta_{T^*} > \alpha_{X^*}$ because otherwise no truncation occurs. We will comment at the end of the section on estimation when the assumptions are not valid.

Now let us present several useful probability formulas for the LT model. The joint distribution function of the observed random variables (T, Y) is

$$F^{T,X}(t,x) = F^{T^*,X^*|T^* \leq X^*}(t,x) = p^{-1}\mathbb{P}(T^* \leq t, X^* \leq x, 0 \leq T^* \leq X^* \leq x)$$

$$= p^{-1}\int_0^t f^{T^*}(\tau)[\int_\tau^{\max(x,\tau)} f^{X^*}(u)du]d\tau, \tag{6.6.3}$$

where

$$p := \mathbb{P}(T^* \leq X^*) = \int_0^\infty f^{T^*}(t)G^{X^*}(t)dt \tag{6.6.4}$$

is the probability of X^* to be observed (not truncated by T^*). Then, by taking partial derivatives, we get the joint density of (T, X),

$$f^{T,X}(t,x) = p^{-1}f^{T^*}(t)f^{X^*}(x)I(0 \leq t \leq x < \infty). \tag{6.6.5}$$

In its turn, (6.6.5) yields two marginal densities for T and X,

$$f^T(t) = p^{-1} f^{T^*}(t) G^{X^*}(t) = \frac{f^{T^*}(t) g(t)}{F^{T^*}(t)}, \tag{6.6.6}$$

and

$$f^X(x) = p^{-1} f^{X^*}(x) F^{T^*}(x) = \frac{f^{X^*}(x) g(x)}{G^{X^*}(x)}, \tag{6.6.7}$$

where

$$g(u) := \mathbb{P}(T \le u \le X) = p^{-1} F^{T^*}(u) G^{X^*}(u). \tag{6.6.8}$$

Now we are in a position to explain proposed estimators. We begin with estimators for the boundary points of the supports,

$$\hat\alpha_{X^*} := X_{(1)} := \min(X_1, \ldots, X_n), \quad \hat\beta_{T^*} := T_{(n)} := \max(T_1, \ldots, T_n),$$

$$\hat\beta_{X^*} := X_{(n)} := \max(X_1, \ldots, X_n). \tag{6.6.9}$$

Next we estimate the cumulative hazard $H^{X^*}(x) := \int_0^x h^{X^*}(u)du$ for $x < \beta_{X^*}$. Using (6.6.7) and (6.6.8) we may write,

$$H^{X^*}(x) = I(x > \alpha_{X^*}) \int_0^x h^{X^*}(u) du = I(x > \alpha_{X^*}) \int_0^x [f^{X^*}(u)/G^{X^*}(u)] du$$

$$= I(x > \alpha_{X^*}) \int_0^x [f^X(u)/g(u)] du = I(x > \alpha_{X^*}) \mathbb{E}\{I(X \le x) g^{-1}(X)\}. \tag{6.6.10}$$

Note that the cumulative hazard is written as the expectation of a function of observed variables. Hence we can estimate it by a plug-in sample mean estimator,

$$\hat H^{X^*}(x) = n^{-1} \sum_{l=1}^n \frac{I(X_l \le x)}{\hat g(X_l)}, \tag{6.6.11}$$

where

$$\hat g(x) := n^{-1} \sum_{l=1}^n I(T_l \le x \le X_l), \tag{6.6.12}$$

and this is a sample mean estimator of $g(x) = \mathbb{E}\{I(T \le x \le X)\}$. Note that estimator (6.6.11) is zero for $x \le X_{(1)}$ and it is equal to $\hat H^{X^*}(X_{(n)})$ for all $x \ge X_{(n)}$. In what follows we may refer to estimator (6.6.11) as an empirical cumulative hazard.

Now we can explain how to estimate the survival function

$$G^{X^*}(x) := \mathbb{P}(X^* > x) = e^{-\int_0^x h^{X^*}(u)du} = e^{-H^{X^*}(x)}. \tag{6.6.13}$$

We plug the empirical cumulative hazard (6.6.11) in the right side of (6.6.13) and get

$$\hat G^{X^*}(x) := e^{-\hat H^{X^*}(x)}. \tag{6.6.14}$$

The attractive feature of the estimator (6.6.14) is that its construction is simple and it is easy for statistical analysis. For instance, suppose that its asymptotic (as the sample size increases) variance is of interest. Then the asymptotic variance of the cumulative hazard is calculated straightforwardly because it is a sample mean estimator, and then the variance of the survival function is evaluated by the delta method. Let us follow these two steps and calculate the asymptotic variance. First, under the above-made assumptions the asymptotic variance of the empirical cumulative hazard is

$$\lim_{n\to\infty}[n\mathbb{V}(\hat{H}^{X^*}(x))] = \mathbb{E}\{I(X \le x)/g^2(X)\} - [H^{X^*}(x)]^2. \tag{6.6.15}$$

Second, the delta method yields

$$\lim_{n\to\infty}[n\mathbb{V}(\hat{G}^{X^*}(x))] = [G^{X^*}(x)]^2(\mathbb{E}\{I(X \le x)/g^2(X)\} - [H^{X^*}(x))]^2)$$

$$= [G^{X^*}(x)]^2\Big(p\int_0^x \frac{f^{X^*}(u)}{F^{T^*}(u)[G^{X^*}(u)]^2}du - [H^{X^*}(x))]^2\Big). \tag{6.6.16}$$

There are two lines in (6.6.16) and both are of interest to us. The top line explains how the variance may be estimated via a combination of plug-in and sample mean techniques. Furthermore, for a sample mean estimator, under a mild assumption, the central limit theorem yields asymptotic normality, then the delta method tells us that the asymptotic normality is preserved by the exponential transformation (6.6.14), and hence we can use this approach for obtaining confidence bands for the estimator (6.6.14). The bottom line in (6.6.16) is important for our understanding conditions implying consistent estimation of the survival function $G^{X^*}(x)$. For instance, if $\alpha_{T^*} < \alpha_{X^*}$, then the integral in (6.6.16) is finite. To see this, note that $0/0 = 0$ and hence the integral over $u \in [0, x]$, $x > \alpha_{X^*}$ is the same as the integral over $u \in [\alpha_{X^*}, x]$ where $F^{T^*}(u) \ge F^{T^*}(\alpha_{X^*}) > 0$ due to the assumed $\alpha_{T^*} < \alpha_{X^*}$. The situation changes if $\alpha_{T^*} = \alpha_{X^*}$, and this is a case in many applications. Depending on the ratio $f^{X^*}(u)/F^{T^*}(u)$ for u near α_{T^*}, the integral in (6.6.16) is either finite or infinity. This is what makes the effect of the LT on estimation so unpredictable because it may preclude us from consistent estimation of G^{X^*} even if $\alpha_{T^*} = \alpha_{X^*}$. Recall that we made a similar conclusion in Section 6.3 for estimation of the hazard rate. Furthermore, if $\alpha_{T^*} > \alpha_{X^*}$ then no consistent estimation of the distribution of X^* is possible, and it will be explained shortly what may be estimated in this case.

Now we are developing an E-estimator of the probability density $f^{X^*}(x)$. According to (6.6.7) we have the following relation,

$$f^{X^*}(x) = f^X(x)G^{X^*}(x)/g(x) \quad \text{whenever} \quad g(x) > 0. \tag{6.6.17}$$

Suppose that we are interested in estimation of the density over an interval $[a, a+b]$ such that $g(x)$ is positive on the interval. Denote by $\{\psi_j(x)\}$ the cosine basis on $[a, a+b]$. Then (6.6.17) allows us to write for a Fourier coefficient,

$$\theta_j := \int_a^{a+b} f^{X^*}(x)\psi_j(x)dx = \mathbb{E}\{I(X \in [a, a+b])\psi_j(X_l)G^{X^*}(X)/g(X)\}. \tag{6.6.18}$$

The expectation in (6.6.18) implies a plug-in sample mean Fourier estimator

$$\hat{\theta}_j := n^{-1}\sum_{l=1}^n I(X_l \in [a, a+b])\psi_j(X_l)\hat{G}^{X^*}(X_l)/\hat{g}(X_l), \tag{6.6.19}$$

where $\hat{G}^{X^*}$ and $\hat{g}$ are defined in (6.6.14) and (6.6.12), respectively.

In its turn, the Fourier estimator yields a density E-estimator $\hat{f}^{X^*}(x)$, $x \in [a, a+b]$. Further, the corresponding coefficient of difficulty and its estimator are

$$d(a, a+b) = b^{-1}\mathbb{E}\{[G^{X^*}(X)/g(X)]^2\}, \tag{6.6.20}$$

and

$$\hat{d}(a, a+b) := b^{-1}n^{-1}\sum_{l=1}^n [\hat{G}^{X^*}(X_l)/\hat{g}(X_l)]^2. \tag{6.6.21}$$

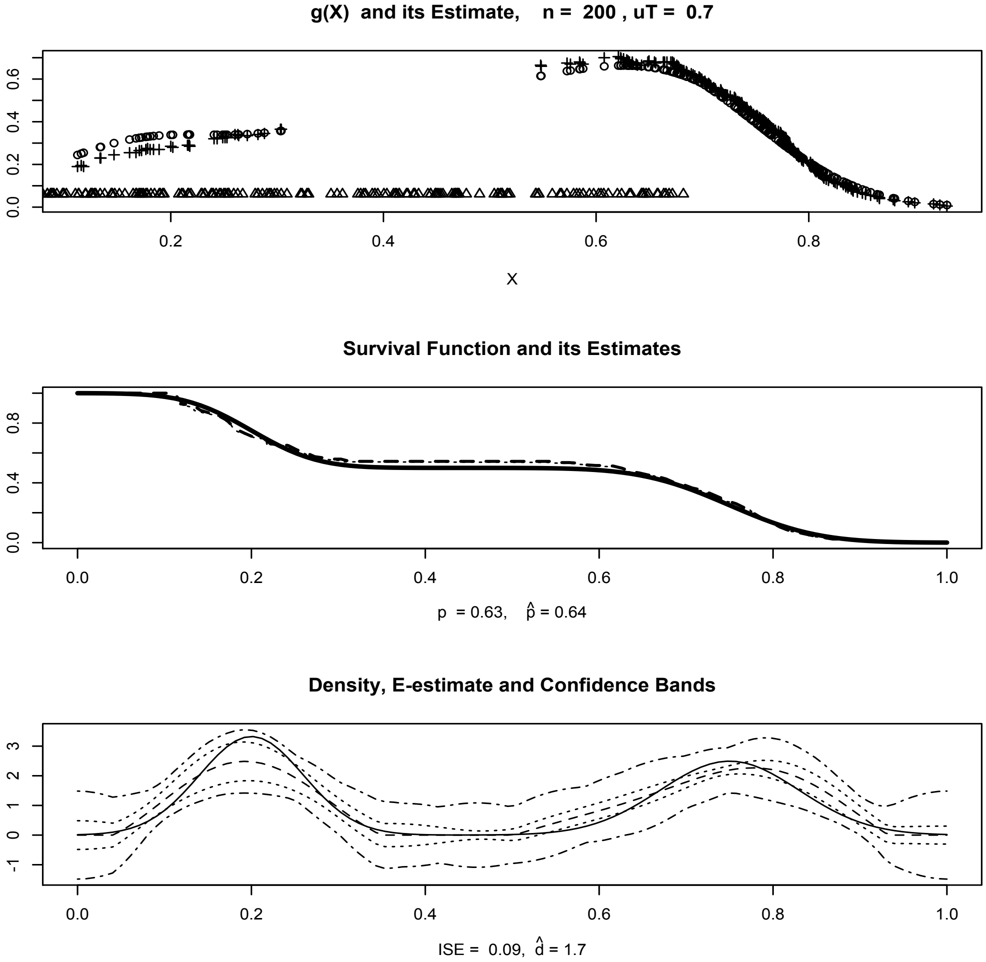


Figure 6.8 *Estimation of the distribution of the lifetime of interest X^* in an LT sample. In the hidden model X^* is distributed according to the Strata (the choice is controlled by parameter corn) and T^* is Uniform($[0, u_T]$). The top diagram shows by circles and crosses $g(X_l)$ and $\hat{g}(X_l)$, and x-coordinates of triangles show observations of the truncating variable T. In the middle diagram, the wide solid, dotted and dashed lines show the underlying survival function, Kaplan–Meyer estimate and sample mean estimate (6.6.14). The subtitle shows the underlying probability (6.6.4) and its estimate (6.6.22). The bottom diagram shows the underlying density (the solid line), the E-estimate (the dashed line), and $(1 - \alpha)$ pointwise (the dotted lines) and simultaneous (the dot-dashed lines) confidence bands. [n = 200, corn = 4, uT = 0.7, alpha = 0.05, cJ0 = 4, cJ1 = 0.5, cTH = 4]*

There is one more interesting unknown parameter that may be estimated. Recall that $p := \mathbb{P}(T^* \leq X^*)$ defines the probability of a hidden pair (T^*, X^*) to be observed, or in other words p defines the likelihood of X^* to be not truncated. Of course it is of interest to know this parameter of the LT. Equality (6.6.8), together with $F^{T^*}(\beta_{T^*}) = 1$, implies that $g(\beta_{T^*}) = p^{-1}G^{X^*}(\beta_{T^*})$. This motivates the following estimator,

$$\hat{p} := \hat{G}^{X^*}(T_{(n)})/\hat{g}(T_{(n)}). \tag{6.6.22}$$

Note that any $t \geq T_{(n)}$ can be also used in (6.6.22) in place of $T_{(n)}$.

Figure 6.8 allows us to look at a simulated LT sample and evaluate performance of the proposed estimators, explanation of its three diagrams can be found in the caption. The top diagram allows us to visualize realizations of X and T. Here we also can compare the underlying function $g(x)$ and its sample mean estimate (6.6.12). Note how fast the right tail of $g(x)$ vanishes. Nonetheless, as we shall see shortly, this does not preclude us from estimation of the distribution of X^*. The middle diagram indicates that the sample mean and the Kaplan–Meyer estimators perform similarly. Its subtitle shows the underlying parameter p and its estimate (6.6.22). As we see, we can get a relatively fair idea about a hidden sampling which is governed by a negative binomial distribution with parameters (n, p). Finally, the bottom diagram shows us the E-estimate of the density of X^*. As we know, the Strata is a difficult density to estimate even for the case of direct observations. Here we are dealing with LT observations that are, as we know, biased. Overall the particular outcome is good because we clearly observe the two strata. The confidence bands are also reasonable, and the ISE, shown in the subtitle, is relatively small. Further, the subtitle shows us the estimated coefficient of difficulty (6.6.21).

Now let us return to the middle diagram in Figure 6.8 where the survival function and its estimates are shown. In many applications the right tail of the survival function is of a special interest. Let us explore a new idea of the tail estimation. Note that $F^{T^*}(x) = 1$ whenever $x \geq \beta_{T^*}$. This allows us to write for $x \geq \beta_{T^*}$,

$$\begin{aligned} G^X(x) &= \mathbb{P}(X^* > x | T^* \leq X^*) \\ &= p^{-1}\mathbb{P}(X^* > x, T^* \leq X^*) = p^{-1}G^{X^*}(x), \;\; x \geq \beta_{T^*}. \end{aligned} \tag{6.6.23}$$

We conclude that $G^{X^*}(x) = pG^X(x)$, $x \geq \beta_{T^*}$. The latter is easy to understand because no truncation occurs whenever $X^* \geq \beta_{T^*}$. As a result, for $x \geq \beta_{T^*}$ the survival function of the hidden variable of interest is proportional to the survival function of the observed variable. Parameter p is estimated by (6.6.22), and $G^X(x)$ is estimated by the empirical survival function $\hat{G}^X(x) := n^{-1}\sum_{l=1}^n I(X_l > x)$. This gives us an opportunity to propose a new estimator of the tail of $G^{X^*}(x)$,

$$\tilde{G}^X(x) := \hat{p}^{-1}n^{-1}\sum_{l=1}^n I(X_l > x), \quad x > T_{(n)}. \tag{6.6.24}$$

Figure 6.9 allows us to look at the zoomed-in tails of estimates produced by estimators (6.6.14) and (6.6.24). The two diagrams show results of different simulations for the same LT model. The solid, dashed and dotted lines are the underlying survival function and estimates (6.6.14) and (6.6.24). As we see, the top diagram depicts an outcome where the new estimate is better, this is also stressed by the ratio between the empirical integrated squared error (ISE) of the estimate (6.6.14) and the integrated squared error (ISEN) of the new estimate (6.6.24). Note that there are just few large observations, and this is a rather typical situation with tails. The sample size $n = 75$ is relatively small but it is chosen for better visualization of the estimates. The bottom diagram exhibits results for another simulation, and here the estimate (6.6.14) is better. This is a tie, and we may conclude that in general two simulations are not enough to compare estimators. To resolve the issue, we repeat the simulation 300 times and then statistically analyze ratios of the ISEs. Sample mean and sample median of the ratios are shown in the subtitle, and they indicate better performance of the estimator (6.6.24). Of course, the sample size is too small for estimation of the tail, but the method of choosing between two estimators is statistically sound. The reader is advised to repeat Figure 6.9 with different parameters, compare performance of the two estimators, and gain experience in choosing between several estimators.

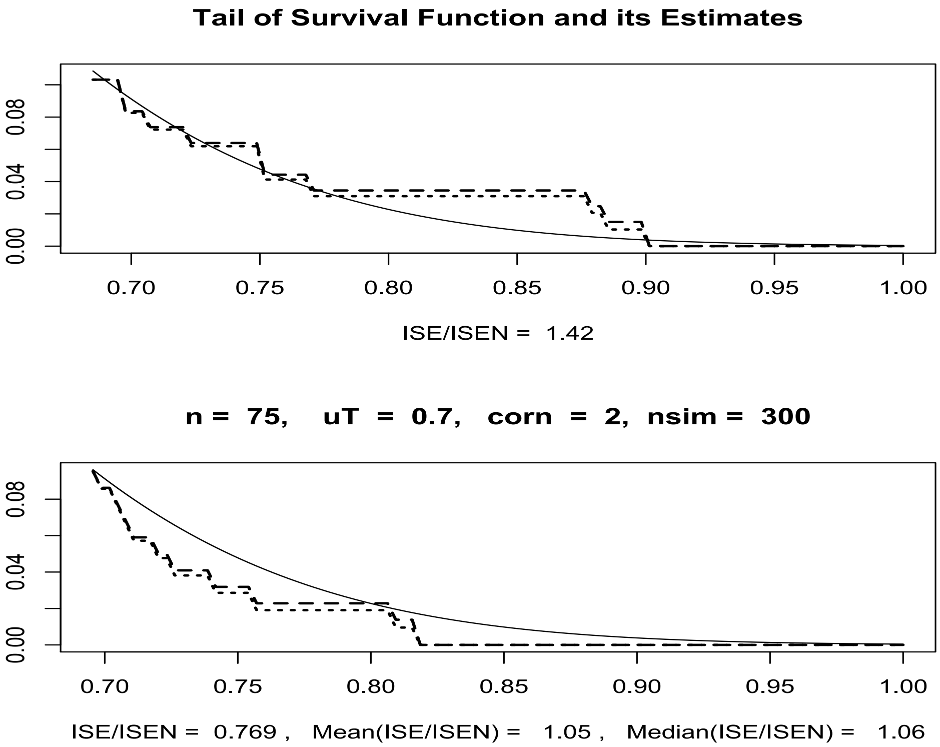


Figure 6.9 *Estimation of the tail of survival function $G^{X^*}(x)$. The underlying simulation is the same as in Figure 6.8. The solid, dashed and dotted lines are the underlying survival function and estimates (6.6.14) and (6.6.24), respectively. {Argument nsim controls the number of simulations.} [n = 75, corn = 2, uT = 0.7, nsim = 300]*

So far we have discussed the problem of estimation of the distribution of a hidden lifetime of interest X^*. It is also of interest to estimate the distribution of a hidden truncating random variable T^*. We again begin with probability formulas. Using (6.6.6) we can write,

$$q^{T^*}(t) := \frac{f^{T^*}(t)}{F^{T^*}(t)} = \frac{f^T(t)}{g(t)} \quad \text{whenever } \ g(t) > 0. \tag{6.6.25}$$

Function $q^{T^*}(t)$ can be estimated using the second equality in (6.6.25), and hence we need to understand how the distribution of interest can be expressed via the function $q(t)$. This function resembles the hazard rate only now the denominator is the cumulative distribution function instead of the survival function. A straightforward algebra implies that

$$F^{T^*}(t) = \exp\Big(- \int_t^{\beta_{T^*}} q^{T^*}(u)du\Big) =: \exp\big(- Q^{T^*}(t)\big)$$

$$= \exp\big(- \mathbb{E}\{I(T > t)/g(T)\}\big), \quad t \in [\alpha_{T^*}, \beta_{T^*}]. \tag{6.6.26}$$

The expectation in (6.6.26) allows us to propose the following plug-in sample mean estimator of the cumulative distribution function (compare with (6.6.14))

$$\hat{F}^{T^*}(t) := \exp\big(- \hat{Q}^{T^*}(t)\big) := \exp\Big(- n^{-1} \sum_{l=1}^{n} I(T_l > t)/\hat{g}(T_l)\Big), \tag{6.6.27}$$

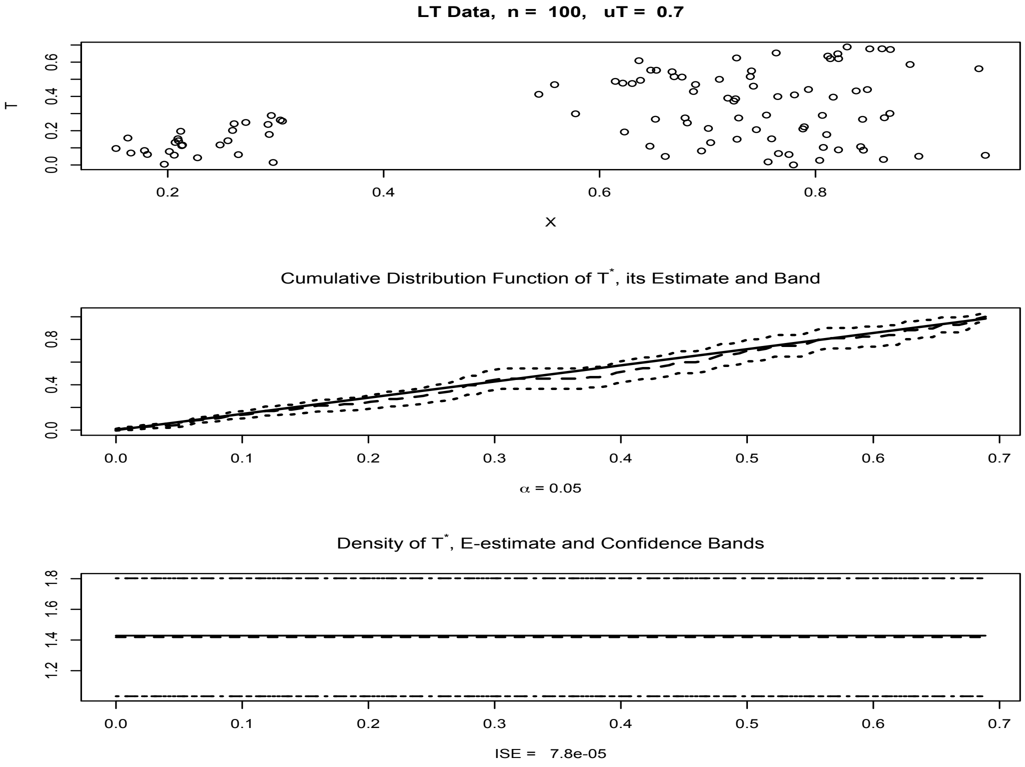


Figure 6.10 *Estimation of the distribution of T^* for LT data. Underlying simulation is the same as in Figure 6.8. Observations are shown in the top diagram. In the middle diagram the solid and dashed lines are the cumulative distribution function F^{T^*} and its estimate. The dotted lines show the $(1-\alpha)$ pointwise confidence band. Curves in the bottom diagram are the same as in the bottom diagram of Figure 6.8. [n = 100, corn = 4, uT = 0.7, alpha = 0.05, cJ0 = 4, cJ1 = 0.5, cTH = 4]*

where $\hat{g}(t)$ is defined in (6.6.12).

Now let us propose an E-estimator of the density $f^{T^*}(t)$ on an interval $[a, a+b]$ such that $g(t)$ is positive on this interval. We are using our traditional notation $\psi_j(t)$ for elements of the cosine basis on $[a, a+b]$. Formula

$$f^{T^*}(t) = f^T(t)F^{T^*}(t)/g(t) \tag{6.6.28}$$

for the density of interest implies that its Fourier coefficients $\kappa_j := \int_a^{a+b} \psi_j(t) f^{T^*}(t)dt$ can be written as

$$\kappa_j = \int_a^{a+b} \frac{f^T(t)\psi_j(t)F^{T^*}(t)}{g(t)}dt = \mathbb{E}\Big\{\frac{I(T \in [a, a+b]\psi_j(T)F^{T^*}(T)}{g(T)}\Big\}. \tag{6.6.29}$$

The expectation in (6.6.29), together with (6.6.27), yields a plug-in sample mean Fourier estimator (compare with (6.6.19))

$$\hat{\kappa}_j := n^{-1} \sum_{l=1}^{n} I(T_l \in [a, a+b])\psi_j(T_l)\hat{F}^{T^*}(T_l)/\hat{g}(T_l). \tag{6.6.30}$$

The Fourier estimator yields a density E-estimator $\hat{f}^{T^*}(t)$, $t \in [a, a+b]$ with the corresponding coefficient of difficulty

$$d(a, a+b) = b^{-1}\mathbb{E}\big\{I(T \in [a, a+b][F^{T^*}(T)/g(T)]^2\big\}. \tag{6.6.31}$$

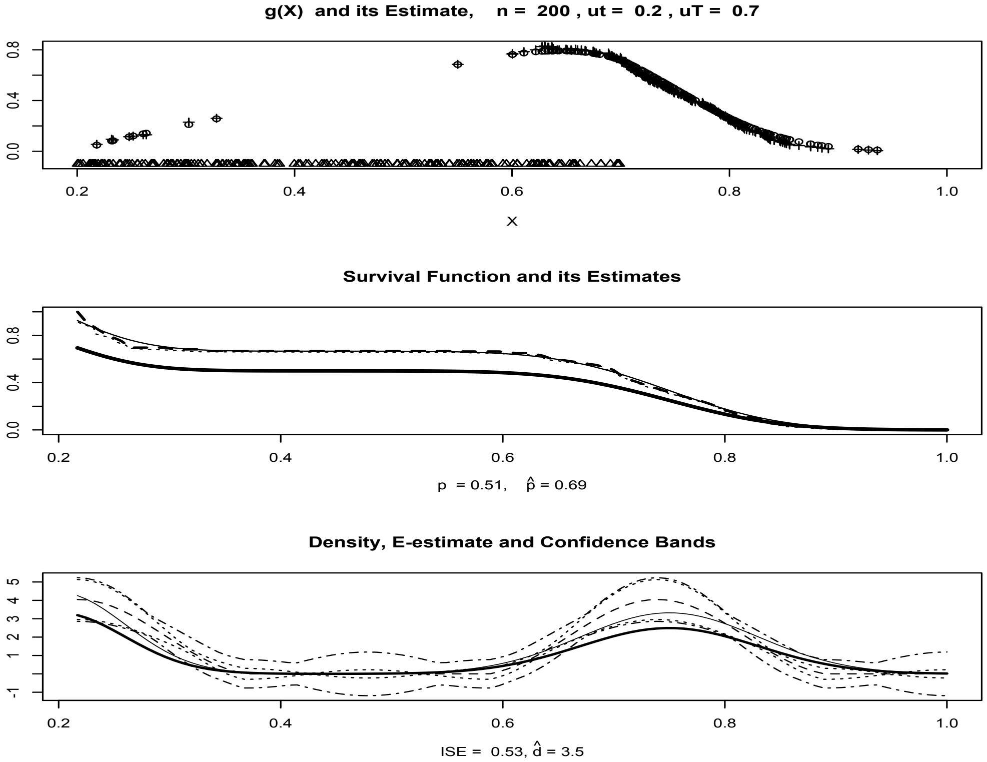


Figure 6.11 *Estimation of the distribution of* X^* *for the case* $\alpha_{X^*} < \alpha_{T^*}$. *This figure is similar to Figure 6.8 apart of two modifications. First, here* T^* *is uniform on interval* $[u_t, u_T] = [0.2, 0.7]$. *Second, the narrow solid lines in the middle and bottom diagrams show the underlying conditional survival function* $G^{X^*|X^*>u_t}(x)$ *and the underlying conditional density* $f^{X^*|X^*>u_t}(x)$*, respectively. [n = 200, corn = 4, ut = 0.2, uT = 0.7, alpha = 0.05, cJ0 = 4, cJ1 = 0.5, cTH = 4]*

Further, typically $[T_{(1)}, T_{(n)})]$ may be recommended as the interval of estimation, and note that $\mathbb{P}(T_{(n)} \leq X_{(n)}) = 1$.

Figure 6.10 shows us how the proposed estimators of the cumulative distribution function and the density of the hidden truncation variable T^* perform. For the particular simulation, the estimates are good. Confidence bands for the density may look too wide, but this is only because the E-estimate is good. The reader is advised to repeat Figure 6.10 with different parameters and test performance of the estimator and confidence bands.

Note that we have developed estimators for T^* from scratch. It was explained in the previous sections that for RC data there is a symmetry between estimation of distributions of the lifetime of interest X^* and censoring variable C^*, and if an estimator for X^* is proposed, then it also can be used for C^*. Is there any type of a similar symmetry for LT data? The answer is "yes." Let γ be a positive constant such that $\gamma \geq \max(\beta_{X^*}, \beta_{T^*})$, and introduce two new random variables $X' := \gamma - T^*$ and $T' := \gamma - X^*$. Then (X', T') can be considered as underlying hidden variables for a corresponding LT data with X' being the lifetime of interest and T' being the truncating variable. This is a type of symmetry that allows us to estimate distributions of T^* using estimators developed for X^*.

Finally, let us explain what may be expected when assumptions (6.6.1) and (6.6.2) are violated. If $\alpha_{X^*} < \alpha_{T^*}$ then the LT hides left tail of the distribution of X^*, and we cannot restore it. Violation of (6.6.2) hides right tail of the distribution of T^*.

Figure 6.11 illustrates the case when $\alpha_{X^*} < \alpha_{T^*}$ (note that the diagrams are similar to the ones in Figure 6.8 whose caption explains the diagrams). First of all, let us look at the top diagram which shows realizations of X and T. If we compare them with those in Figure 6.8, then we may conclude that visualization of observations is unlikely to point upon the violation of assumption (6.6.1). The reader is advised to repeat Figures 6.8 and 6.11 and gain experience in assessing LT data. Further, it is clear from the two bottom diagrams that the estimators are inconsistent, and this is what was predicted. Nonetheless, it looks like the estimates mimic the underlying ones. Let us explore this issue and shed light on what the estimators do when $\alpha_{X^*} < \alpha_{T^*}$.

We begin with the estimator (6.6.14) of the survival function. Recall that we are considering the case $\alpha_{X^*} < \alpha_{T^*}$. Note that then $\mathbb{P}(X \leq \alpha_{T^*}) = 0$, and using (6.6.11) we can write for $x > \alpha_{T^*}$,

$$\mathbb{E}\{\hat{H}^{X^*}(x)\} = \mathbb{E}\Big\{\frac{I(X < x)I(X > \alpha_{T^*})}{\hat{g}(X)}\Big\}$$

$$= \mathbb{E}\Big\{\frac{I(\alpha_{T^*} < X < x)}{g(X)}\Big\} + \mathbb{E}\Big\{I(\alpha_{T^*} < X < x)\Big[\frac{1}{\hat{g}(X)} - \frac{1}{g(X)}\Big]\Big\}. \qquad (6.6.32)$$

The first term in (6.6.32) is what we are interested in because the second one, under a mild assumption, vanishes as n increases. Using (6.6.7) we can express the first term via the hazard rate of X^*,

$$\mathbb{E}\Big\{\frac{I(\alpha_{T^*} < X < x)}{g(X)}\Big\} = \int_{\alpha_{T^*}}^{x} h^{X^*}(u)du. \qquad (6.6.33)$$

In its turn, (6.6.33) implies the following relation,

$$e^{-\mathbb{E}\big\{\frac{I(\alpha_{T*} < X < x)}{g(X)}\big\}} = e^{-\int_{\alpha_{T^*}}^{x} h^{X^*}(u)du}$$

$$= \frac{e^{-\int_0^x h^{X^*}(u)du}}{e^{-\int_0^{\alpha_{T^*}} h^{X^*}(u)du}} = \frac{G^{X^*}(x)}{G^{X^*}(\alpha_{T^*})} =: G^{X^*|X^* > \alpha_{T^*}}(x). \qquad (6.6.34)$$

On the right side of (6.6.34) we see the conditional survival function of X^* given $X^* > \alpha_{T^*}$, and this is the characteristic that estimator (6.6.14) estimates.

We may conclude that it is more accurate to say that the survival function estimator (as well as the Kaplan–Meier estimator) estimates the conditional density $G^{X^*|X^* > \alpha_{T^*}}(x)$ because $G^{X^*|X^* > \alpha_{T^*}}(x) = G^{X^*}(x)$ whenever $\alpha_{X^*} \geq \alpha_{T^*}$. This is a nice conclusion because, regardless of the assumption (6.6.1), we estimate a meaningful characteristic of the random variable of interest.

If we return to the middle diagram in Figure 6.11, then we can see that the survival function estimate is above the underlying survival function (the wide solid line), and now we know why. Further, the narrow solid line exhibits the underlying conditional survival function $G^{X^*|X^* > \alpha_{T^*}}(x)$, and we may note that the E-estimator does a good job in estimation of the conditional survival function. Further, note that the estimated probability $\hat{p} = 0.69$ of avoiding the truncation is larger than the underlying $p = 0.53$. To understand why we need to look at the estimator (6.6.22) and recall that for the considered setting the estimator $\hat{G}^{X^*}(T_{(n)})$ estimates the conditional density $G^{X^*|X^* > \alpha_{T^*}}(\beta_T) = G^{X^*}(\beta_T)/G^{X^*}(\alpha_{T^*})$. As a result, we can expect that the estimate $\hat{p}$ will be larger than the underlying p by the factor $1/G^{X^*}(\alpha_{T^*})$. And indeed, $\hat{p}/p = 1.30$ and this is fairly close to $1/G^{X^*}(\alpha_{T^*}) = 1.33$.

Now let us turn our attention to the bottom diagram in Figure 6.11 which is devoted to estimation of the density $f^{X^*}(x)$. The estimate (the dashed line) is far from the underlying density (the wide solid line), and this supports the above-made conclusion that if (6.6.1) is violated then consistent estimation of the density is impossible. Nonetheless, we may notice that the estimate mimics the density's shape. As a result, let us explore the density estimator for the case when (6.6.1) does not hold.

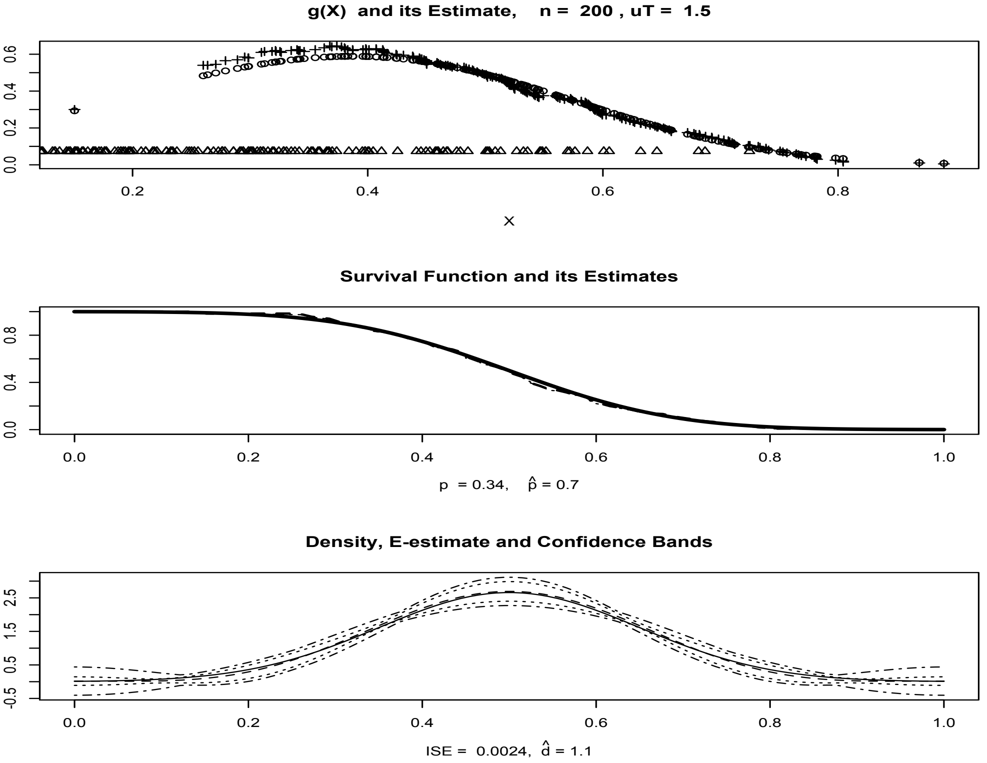


Figure 6.12 *Estimation of the distribution of* X^* *for the case* $\beta_{T^*} > \beta_{X^*}$. *The figure is created by Figure 6.8 by setting* $uT = 1.5$ *and* $corn = 2$.

Consider the expectation of Fourier estimator (6.6.19) given $\alpha_{X^*} < \alpha_{T^*}$. Let $[a, a+b]$ be an interval such that $g(x) > 0$ for $x \in [a, a+b]$. We may write using (6.6.7),

$$\mathbb{E}\{\hat{\theta}_j\} = \mathbb{E}\{I(X \in [a, a+b])\psi_j(X)\hat{G}^{X^*}(X)/\hat{g}(X)\}$$

$$= \mathbb{E}\{I(X \in [a, a+b])\psi_j(X)G^{X^*|X^*>\alpha_{T^*}}(X)/g(X)\}$$

$$+\mathbb{E}\Big\{I(X \in [a, a+b])\psi_j(X)\big[\frac{\hat{G}^{X^*}(X)}{\hat{g}(X)} - \frac{G^{X^*|X^*>\alpha_{T^*}}(X)}{g(X)}\big]\Big\}. \quad (6.6.35)$$

The first expectation on the right side of (6.6.35) is the term of interest because the second one, under a mild assumption, vanishes as n increases. Using (6.6.17) and (6.6.34) we may write,

$$\mathbb{E}\{I(X \in [a, a+b])\psi_j(X)G^{X^*|X^*>\alpha_{T^*}}(X)/g(X)\}$$

$$= \int_a^{a+b} f^X(x)\psi_j(x)G^{X^*|X^*>\alpha_{T^*}}(x)[g(x)]^{-1}dx = \int_a^{a+b} \frac{f^{X^*}(x)\psi_j(x)}{G^{X^*}(\alpha_{T^*})}dx. \quad (6.6.36)$$

Introduce the conditional density,

$$f^{X^*|X^*>\alpha_{T^*}}(x) := \frac{f^{X^*}(x)}{G^{X^*}(\alpha_{T^*})}I(x > \alpha_{T^*}). \quad (6.6.37)$$

Combining (6.6.35)-(6.6.37) we conclude that the Fourier estimator (6.6.19) estimates Fourier coefficients of the conditional density (6.6.37).

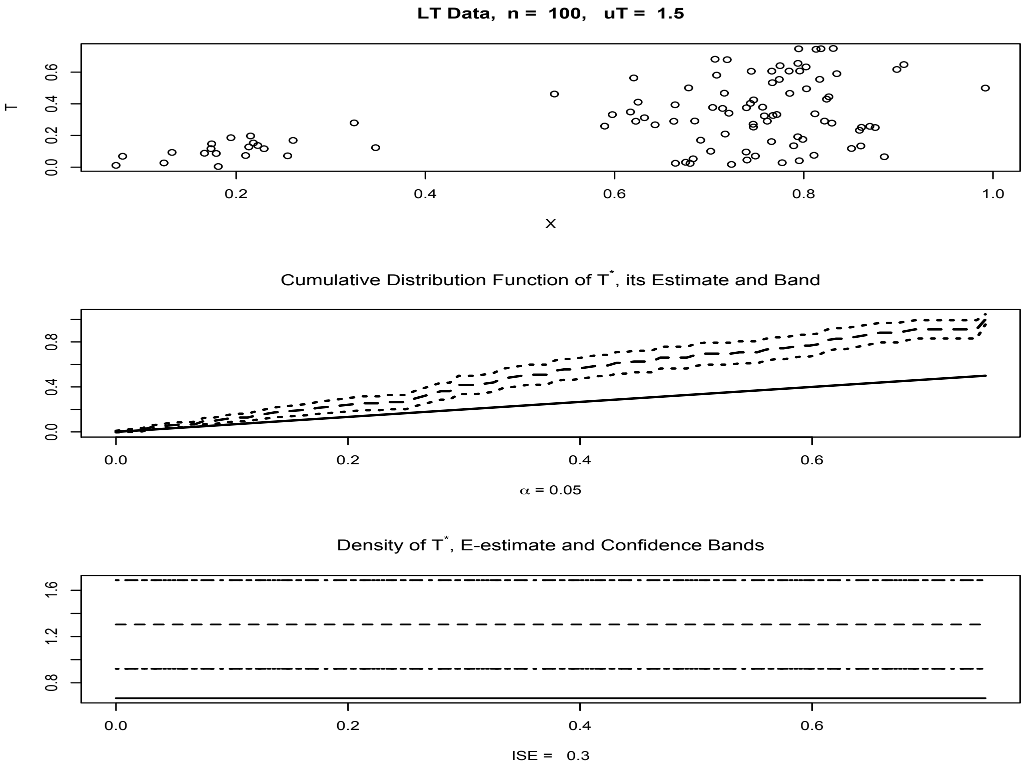


Figure 6.13 *Estimation of the distribution of T^* for the case $\beta_{T^*} > \beta_{X^*}$. The figure is created by Figure 6.10 via setting $uT = 1.5$.*

If we now return to the bottom diagram in Figure 6.11, the narrow solid line shows us the underlying conditional density. The E-estimate (the dashed line) is far from being perfect but its shape correctly indicates the two strata in the underlying conditional density.

Now let us consider the case $\beta_{T^*} > \beta_{X^*}$. Figures 6.12 and 6.13 shed light on this case. Figure 6.12 indicates that this case does not preclude us from consistent estimation of the distribution of X^*. In the top diagram x-coordinates of circles and triangles show us the observed values of X and T, respectively. Our aim is to explore the possibility to realize that $\beta_{T^*} > \beta_{X^*}$. As we discussed earlier, theoretically $T_{(n)}$ should approach $\beta_T = \beta_{X^*} = 1$. Because $X_{(n)}$ converges in probability to $\beta_X = 1$, it may be expected that $T_{(n)}$ and $X_{(n)}$ are close to each other. Unfortunately, even for the used relatively large sample size $n = 200$, $T_{(n)}$ is significantly smaller than $X_{(n)}$ and the outcome resembles what we observed in Figure 6.8. The explanation of this observation is based on formula (6.6.6) for the density $f^T(t)$. It indicates that the density is proportional to the survival function $G^{X^*}(t)$ which vanishes (and in our case relatively fast) as $t \to \beta_X = 1$. This is what significantly slows down the convergence of $T_{(n)}$ to $X_{(n)}$. The lesson learned is that data alone may not allow us to verify validity of (6.6.2). Further, we cannot consistently estimate p. At the same time, as it could be expected, consistent estimation of the distribution of X^* is possible and the exhibited results support this conclusion.

The situation clearly changes when the aim is to estimate the distribution of T^*. Figure 6.13 shows a particular outcome, and it clearly indicates our inability to estimate the distribution of T^*. At the same time, as it follows from formulas, the shape of density $f^{T^*}(t)$ over interval $[T_{(1)}, T_{(n)}]$ may be visualized, and the bottom diagram sheds light on this conclusion. The interested reader may theoretically establish what the proposed estimator estimates for the considered case $\beta_{T^*} > \beta_{X^*}$.

The reader is advised to repeat Figures 6.8–6.13 with different parameters, pay a special attention to feasible intervals of estimation, and get used to statistical analysis of LT data.

6.7 Estimation of Distributions for LTRC Data

As we already know from Section 6.4, in many applications the LT and the RC occur together, and then we are dealing with left truncated and right censored (LTRC) data. A nice example of LTRC, discussed in Section 6.4, is the actuarial example where the random variable of interest is the loss which is left truncated and right censored by the policy's deductible and limit on payment, respectively. Let us also recall several teachable moments learned in that section. First, it is important to realize that data are modified. Second, LTRC may preclude us from consistent estimation. In particular, LT may make impossible estimation of the left tail of the distribution of an underlying lifetime of interest while the RC may make impossible estimation of its right tail. Third, it is important to check assumptions about underlying distributions. In particular, it may be possible to estimate the distribution of the lifetime of interest but not the distribution of the censoring variable and vise versa. Finally, choosing a feasible interval of estimation becomes a critical part of an estimation procedure.

The aim of this section is to explore the problem of estimation of survival functions and probability densities for variables hidden by LTRC modification.

Following Section 6.4, let us briefly recall the LTRC mechanism of generating a sample of size n. There is a hidden sequential sampling from a triplet of nonnegative random variables (T^*, X^*, C^*) whose joint distribution is unknown. T^* is the truncation random variable, X^* is the random variable (lifetime) of interest, and C^* is the censoring random variable. Suppose that (T_k^*, X_k^*, C_k^*) is the kth realization of the hidden triplet and before this realization a sample of size $l-1$, $l-1 \leq \min(k-1, n-1)$ of LTRC observations is collected. Then if $T_k^* > \min(X_k^*, C_k^*)$ then left truncation of the kth realization occurs meaning that: (i) The kth triplet is not observed; (ii) The fact that the kth observation occurred is unknown; (iii) Next realization of the triplet occurs. On the other hand, if $T_k^* \leq \min(X_k^*, C_k^*)$ then the observation $(T_l, V_l, \Delta_l) := (T_k^*, \min(X_k^*, C_k^*), I(X_k^* \leq C_k^*))$ is added to the observed LTRC sample. The hidden sequential sampling from the triplet (T^*, X^*, C^*) stops as soon as $l = n$.

Now let us present basic probability formulas. The probability of observing a realization of the hidden triplet is

$$p := \mathbb{P}(T^* \leq \min(X^*, C^*)). \tag{6.7.1}$$

The joint cumulative distribution function of the observed triplet of random variables is

$$\begin{aligned} F^{T,V,\Delta}(t, v, \delta) &:= \mathbb{P}(T \leq t, V \leq v, \Delta \leq \delta) \\ &= p^{-1}\mathbb{P}(T^* \leq t, T^* \leq V^* \leq v, \Delta^* \leq \delta). \end{aligned} \tag{6.7.2}$$

If we additionally assume that hidden random variables T^*, X^* and C^* are independent and continuous, then (6.7.2) yields the following joint mixed density,

$$f^{T,V,\Delta}(t, v, \delta) = p^{-1} f^{T^*}(t) I(t \leq v) \Big[f^{X^*}(v) G^{C^*}(v)\Big]^{\delta} \Big[f^{C^*}(v) G^{X^*}(v)\Big]^{1-\delta}. \tag{6.7.3}$$

In its turn, (6.7.3) yields a marginal density

$$f^{V,\Delta}(v, 1) = p^{-1} f^{X^*}(v) G^{C^*}(v) F^{T^*}(v) = h^{X^*}(v)[p^{-1} G^{C^*}(v) F^{T^*}(v) G^{X^*}(v)], \tag{6.7.4}$$

where $h^{X^*}(x) := f^{X^*}(x)/G^{X^*}(x)$ is the hazard rate of X^*. Further, using formula $G^{V^*}(t) = G^{X^*}(t)G^{C^*}(t)$ and (6.7.2) we get

$$f^T(t) = p^{-1} f^{T^*}(t)G^{X^*}(t)G^{C^*}(t). \tag{6.7.5}$$

Further, (6.7.4) yields

$$f^{X^*}(x) = \frac{f^{V,\Delta}(x,1)}{p^{-1}G^{C^*}(x)F^{T^*}(x)} \quad \text{whenever } G^{C^*}(x)F^{T^*}(x) > 0. \tag{6.7.6}$$

Finally, we introduce a probability that plays a key role in the analysis of LTRC data,

$$g(x) := \mathbb{P}(T \le x \le V) = p^{-1}F^{T^*}(x)G^{X^*}(x)G^{C^*}(x). \tag{6.7.7}$$

Now let us formulate assumptions motivated by previous sections. Recall notations α_Z and β_Z for the lower and upper bounds of the support of a variable Z. In what follows it is assumed, in addition to the mutual independence of continuous variables T^*, X^* and C^*, that

$$\min(\alpha_{X^*}, \alpha_{C^*}) \ge \alpha_{T^*}, \quad \beta_{X^*} \ge \beta_{T^*}, \quad \beta_{C^*} \ge \beta_{X^*}. \tag{6.7.8}$$

Now we are in a position to propose estimators for the survival function and density of the lifetime of interest X^*. We begin with estimation of the survival function $G^{X^*}(x)$. Using (6.7.6) and (6.7.7) we conclude that the density of X^* can be written as

$$f^{X^*}(x) = \frac{f^{V,\Delta}(x,1)G^{X^*}(x)}{g(x)}, \quad x \in [0, \beta_{X^*}). \tag{6.7.9}$$

Formula (6.7.9) allows us to obtain a simple formula for the cumulative hazard of X^*,

$$H^{X^*}(x) := \int_0^x [f^{X^*}(u)/G^{X^*}(u)]du$$

$$= \int_0^x [f^{V,\Delta}(u,1)/g(u)]du = \mathbb{E}\{\Delta I(V \le x)g^{-1}(V)\}, \quad x \in [0, \beta_{X^*}). \tag{6.7.10}$$

Recall that $G^{X^*}(x) = \exp\{-H^{X^*}(x)\}$, and then the expectation on the right side of (6.7.10) implies the following plug-in sample mean estimator of the survival function,

$$\hat{G}^{X^*}(x) := \exp\Big\{ - n^{-1} \sum_{l=1}^{n} \frac{\Delta_l I(V_l \le x)}{\hat{g}(V_l)} \Big\}. \tag{6.7.11}$$

Here

$$\hat{g}(x) := n^{-1} \sum_{l=1}^{n} I(T_l \le x \le V_l) \tag{6.7.12}$$

is the sample mean estimator of the probability $g(x)$ defined in (6.7.7). Further, it is a straightforward calculation to find an asymptotic expression for the variance of empirical survival function (6.7.11),

$$\lim_{n\to\infty} n\mathbb{V}(\hat{G}^{X^*}(x)) = [G^{X^*}(x)]^2[\mathbb{E}\{\Delta I(V \le x)[g(V)]^{-2}\} - (H^{X^*}(x))^2]. \tag{6.7.13}$$

This result, together with the central limit theorem and delta method, allows us to get a pointwise confidence band for the empirical survival function.

To use our E-estimation methodology for estimation of the density $f^{X^*}(x)$ over an interval $[a, a+b] \subset [\alpha_{X^*}, \beta_{X^*})$, we need to understand how to express its Fourier coefficients as expectations. Recall our notation $\{\psi_j(x)\}$ for the cosine basis on $[a, a+b]$. We can write with the help of (6.7.6) and (6.7.7) that

$$\theta_j := \int_a^{a+b} \psi_j(x) f^{X^*}(x)dx = \int_a^{a+b} \frac{\psi_j(x) f^{V,\Delta}(x,1) G^{X^*}(x)}{g(x)} dx$$

$$= \mathbb{E}\Big\{ \frac{\Delta I(V \in [a, a+b]) \psi_j(V) G^{X^*}(V)}{g(V)} \Big\}. \qquad (6.7.14)$$

The expectation in (6.7.14) yields the following plug-in sample mean Fourier estimator,

$$\hat{\theta}_j := n^{-1} \sum_{l=1}^{n} \Delta_l I(V_l \in [a, a+b]) \psi_j(V_l) \hat{G}^{X^*}(V_l)/\hat{g}(V_l). \qquad (6.7.15)$$

Fourier estimator (6.7.15) yields a density E-estimator $\hat{f}^{X^*}(x)$ with the coefficient of difficulty

$$d(a, a+b) = b^{-1} \mathbb{E}\{\Delta I(V \in [a, a+b])[G^{X^*}(V)/g(V)]^2\}$$

$$= pb^{-1} \int_a^{a+b} \frac{f^{X^*}(x)}{F^{T^*}(x) G^{C^*}(x)} dx. \qquad (6.7.16)$$

The coefficient of difficulty is of a special interest to us because it sheds light on a feasible interval of estimation. Namely, we know that $F^{T^*}(x)$ vanishes as $x \to \alpha_{T^*}$ and $G^{C^*}(x)$ vanishes as $x \to \beta_{C^*}$, and this is what may make the integral in (6.7.16) large. The made assumption (6.7.8) allows us to avoid the case of infinite coefficient of difficulty by choosing an interval of estimation satisfying $[a, a+b] \subset [V_{(1)}, V_{(n)}]$, but still the coefficient of difficulty may be prohibitively large. At the same time, vanishing tails of an underlying density f^{X^*} may help in keeping the coefficient of difficulty reasonable, while an increasing tail makes estimation more complicated.

Now let us stress that if $\alpha_{T^*} > \alpha_{X^*}$ then, as it was explained in Section 6.6, the proposed estimators estimate the conditional survival function $G^{X^*|X^*>\alpha_{T^*}}(x)$ and the conditional density $f^{X^*|X^*>\alpha_{T^*}}(x)$. Of course, these conditional characteristics coincide with $G^{X^*}(x)$ and $f^{X^*}(x)$ whenever $\alpha_{T^*} \leq \alpha_{X^*}$, and hence we may say that in general we estimate those conditional characteristics. Statistical analysis of this setting is left as an exercise.

Figure 6.14 illustrates performance of the proposed estimators of the conditional survival function and the conditional density of the lifetime of interest. Its caption explains the simulation and diagrams. Note that the assumption (6.7.8) holds for the particular simulation, and choosing parameter $u_t > 0$ allows us to consider the case $\alpha_{T^*} > \alpha_{X^*}$ and then test estimation of the conditional characteristics.

The top diagram in Figure 6.14 shows us simulated LTRC data generated by the Normal lifetime and uniformly distributed truncating and censoring random variables. These three variables are mutually independent. The good news here is that the number $N = 230$ of available uncensored observations of the lifetime of interest is relatively large with respect to the sample size $n = 300$. If we look at the smallest observations, we may observe that α_{T^*} is close to zero, and this implies a chance for a good estimation of the left tail of the distribution of X^*. We may also conclude that it is likely that $\beta_{X^*} < \beta_{C^*}$ because a relatively large number of uncensored observations are available to the right of the largest observation of the censoring variable. The middle diagram shows us the proposed plug-in sample mean estimate (6.7.11) of $G^{X^*}(x)$, the Kaplan–Meier estimate, and the above-explained 95% confidence band for the plug-in sample mean estimate. The two estimates are practically the same. Despite a relatively large sample size $n = 300$, we see a pronounced deviation of

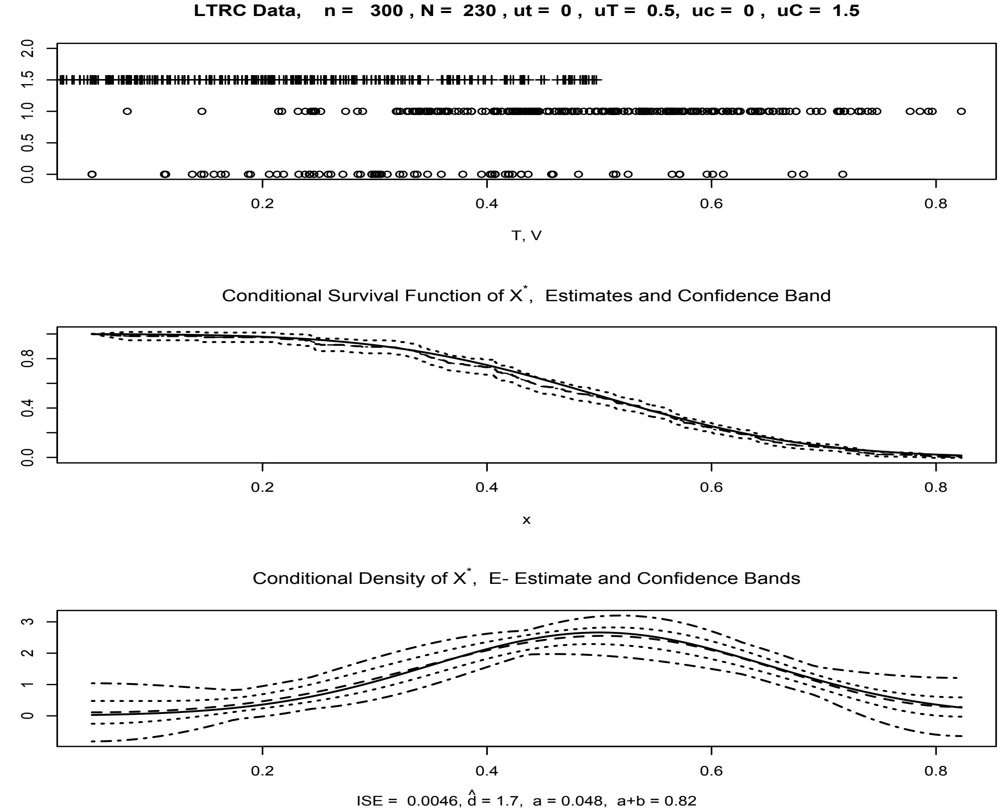


Figure 6.14 *Estimation of the conditional survival function* $G^{X^*|X^*>\alpha_{T^*}}(x)$ *and the conditional density* $f^{X^*|X^*>\alpha_{T^*}}(x)$ *for the case of LTRC observations generated by independent and continuous hidden variables. In the used simulation* T^* *is Uniform*(u_t, u_T)*,* X^* *is the Normal and* C^* *is Uniform*(u_c, u_C)*; the used parameters are shown in the main title. The top diagram shows a sample of size* $n = 300$ *from* (T, V, Δ)*. Observations of* (V, Δ) *are shown by circles,* $N := \sum_{l=1}^{n} \Delta_l = 230$ *is the number of uncensored observations. Observations of the truncation variable* T *are shown via horizontal coordinates of crosses. In the middle diagram, the solid line is the underlying conditional survival function* $G^{X^*|X^*>\alpha_{T^*}}(x)$ *shown for* $x \in [V_{(1)}, V_{(n)}]$*, the dashed and dot-dashed lines are the sample-mean and Kaplan-Meier estimates (they are close to each other), and the dotted lines show the* $(1-\alpha)$ *pointwise confidence band. The bottom diagram shows the underlying conditional density (the solid line), its E-estimate (the dashed line), and the pointwise (dotted lines) and simultaneous (dot-dashed lines)* $1-\alpha$ *confidence bands. The E-estimate is for interval* $[a, a+b]$ *with default values* $a = V_{(1)}$ *and* $a + b = V_{(n)}$ *shown in the subtitle.* $\{$*Distribution of* T^* *is either the Uniform*(u_t, u_T) *or Exponential*(λ_T) *where* λ_T *is the mean. Censoring distribution is either Uniform*(u_c, u_C) *or Exponential*(λ_C)*. Parameters of underlying distributions are shown in the title of the top diagram. To choose, for instance, exponential truncation and censoring, set trunc = "Expon", cens = "Expon" and then either use default parameters or assign wished ones. To choose a manual interval* $[a, a+b]$*, assign wished values to arguments a and b.*$\}$ *[n = 300, corn = 2, trunc = "Unif", ut = 0, uT = 0.5, lambdaT = 0.3, cens = "Unif", uc = 0, uC = 1.5, lambdaC = 1.5, a = NA, b = NA, alpha = 0.05, cJ0 = 4, cJ1 = 0.5, cTH = 4]*

the estimates from the underlying survival function, and we can also see the relatively large width of the band which predicts such a possibility. In the bottom diagram, despite the skewed LTRC observations, the density E-estimate correctly shows the unimodal shape of the underlying Normal density. The estimated coefficient of difficulty, shown in the subtitle, is equal to 1.7. The latter implies that, with respect to the case of direct observations of X^*, we need 70% more LTRC observations to get the same MISE. Figure 6.14 is a good learning tool to explore LRTC data and estimators. It also allows a user to manually choose an interval of estimation $[a, a+b]$, and this will be a valuable lesson on its own.

What can be said about estimation of the distribution of C^* and T^* as well as the parameter p? Recall our discussion in Section 6.5 that X^* and C^* are "symmetric" random variables in the sense that if we consider $1-\Delta$ instead of Δ, then formally C^* becomes the random variable of interest and X^* becomes the censoring random variable. As a result, we can use the proposed estimators for estimation of the distribution of C^*, only by doing so we need to keep in mind the assumptions and then correspondingly choose a feasible interval of estimation. The problem of estimation of the distribution of T^* and parameter p is not new for us because we can consider (T, V) as an LT realization of an underlying pair (T^*, V^*). Then Section 6.6 explains how the distribution of T^* and parameter p can be estimated.

We finish the section by considering a case where the main assumption about mutual independence and continuity of the triplet of hidden random variables is no longer valid. In some applications it is known that the censoring random variable is not smaller than the truncating variable, and the censoring random variable may have a mixed distribution. As an example, we may consider the model of a clinical trial where

$$C^* := T^* + U^* := T^* + [u_C B^* + (1-B^*)U'], \tag{6.7.17}$$

u_C is a positive constant, U' is a nonnegative continuous random variable with the support $[u_c, u_C]$, and B^* is a Bernoulli random variable with $\mathbb{P}(B^*=1) = \mathbb{P}(U^* = u_C)$ being the probability that X^* is censored by the end of a clinical trial.

Assume that X^* is continuous and independent from (T^*, C^*), while truncation and censoring variables may be dependent and have a mixed (continuous and discrete) joint distribution. Can our estimators for the distribution of X^* be used in this case? In other words, are the estimators robust? To answer this question theoretically, we need to understand if formula (6.7.9), which was used to propose the estimators, still holds for the considered setting. We begin with a formula for the probability $g(x)$. Write,

$$\begin{aligned} g(x) &:= \mathbb{P}(T \le x \le V) = \mathbb{P}(T^* \le x \le V^* | T^* \le V^*) \\ &= \frac{\mathbb{P}(T^* \le x \le V^*, T^* \le V^*)}{\mathbb{P}(T^* \le V^*)} = \frac{\mathbb{P}(T^* \le x \le V^*)}{p} \\ = p^{-1}\mathbb{P}(T^* \le x, X^* \ge x, &C^* \ge x) = p^{-1} G^{X^*}(x)\mathbb{P}(T^* \le x \le C^*). \end{aligned} \tag{6.7.18}$$

In the last equality the independence of X^* and (T^*, C^*) was used. Next, we can write,

$$\begin{aligned} \mathbb{P}(V \le x, \Delta = 1) &= \mathbb{P}(X^* \le x, X^* \le C^* | T^* \le V^*) \\ = p^{-1}\mathbb{P}(X^* \le x, T^* \le X^* \le C^*) &= p^{-1}\int_0^x f^{X^*}(u)\mathbb{P}(T^* \le u \le C^*)du. \end{aligned} \tag{6.7.19}$$

Differentiation of (6.7.19) with respect to x and then using (6.7.18) allows us to write,

$$f^{V,\Delta}(x,1) = p^{-1} f^{X^*}(x)\mathbb{P}(T^* \le x \le C^*) = \frac{f^{X^*}(x)g(x)}{G^{X^*}(x)}. \tag{6.7.20}$$

In its turn, (6.7.20) implies that the density of interest can be written as

$$f^{X^*}(x) = \frac{f^{V,\Delta}(x,1)G^{X^*}(x)}{g(x)}, \quad 0 \le x < \beta_{X^*}. \tag{6.7.21}$$

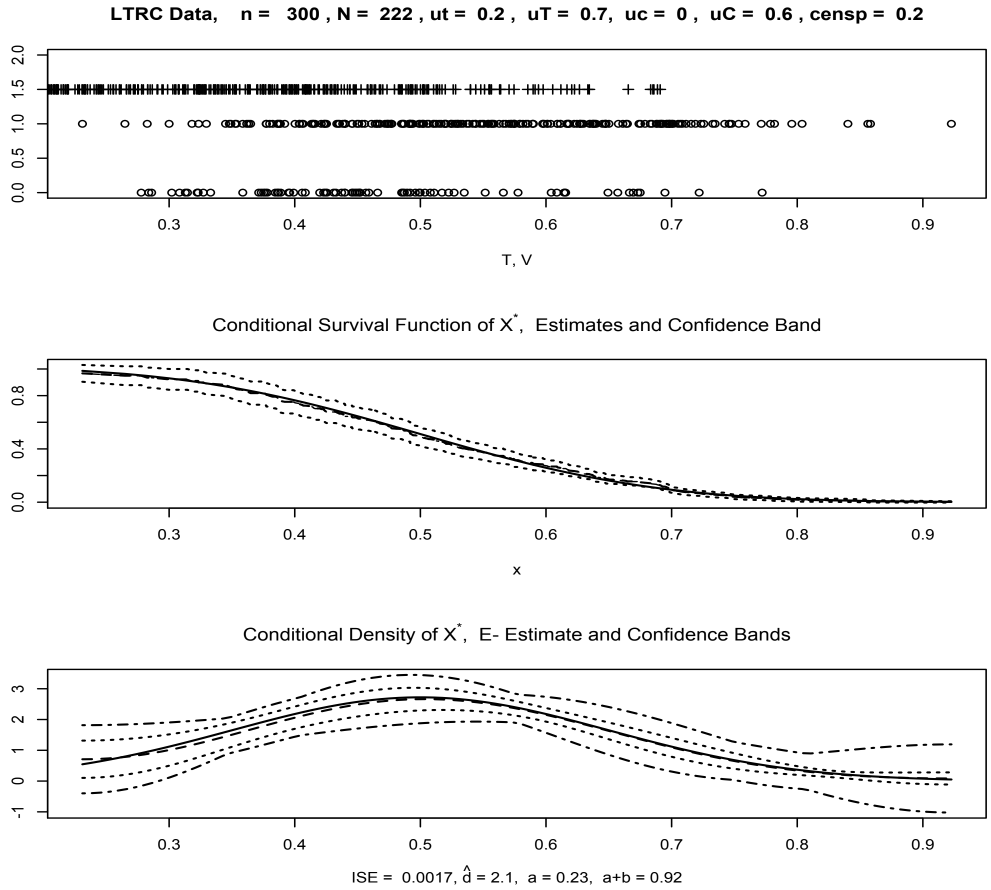


Figure 6.15 *Estimation of the conditional survival function and the conditional density of a lifetime of interest for LTRC data when $C^* := T^* + U^*$ as in (6.7.17). Variables in the triplet (X^*, T^*, U^*) are mutually independent. Variable U^* is the mixture of Uniform(u_c, u_C) random variable with the constant u_C, and the probability $\mathbb{P}(U^* = u_C)$ is controlled by the argument censp. Otherwise the underlying simulation and the structure of diagrams are as in Figure 6.14. [n = 300, corn = 2, trunc = "Unif", ut = 0.2, uT = 0.7, lambdaT = 0.3, cens = "Unif", uc = 0, uC = 0.6, lambdaC = 1.5, censp = 0.2, a = NA, b = NA, alpha = 0.05, cJ0 = 4, cJ1 = 0.5, cTH = 4]*

This is the same formula as (6.7.9) which was established for the case of independent and continuous hidden random variables. Hence an estimator, motivated by that formula, can be used for the studied case as well. Further, if $\alpha_{X^*} > \alpha_{T^*}$ then we estimate conditional characteristics $G^{X^*|X^*>\alpha_{T^*}}(x)$ and $f^{X^*|X^*>\alpha_{T^*}}(x)$.

Figure 6.15 allows us to test the made conclusion for the model (6.7.17). Its structure is similar to Figure 6.14 and the caption explains the underlying LTRC mechanism where the parameter *censp* controls the choice of $\mathbb{P}(U^* = u_C)$. In the considered simulation this parameter is equal to 0.2, meaning that in the example of a clinical trial 20% of participants are right censored by the end of the trial. The used LTRC mechanism creates challenges for the estimators because $\alpha_T^* = u_t = 0.2 > \alpha_{X^*} = 0$. As a result, theoretically we may estimate only the underlying conditional survival function and the conditional density given $X^* > u_t = 0.2$. The top diagram indicates that the truncation variable is separated from zero and it is likely that α_{T^*} is close to 0.2. Further, note that there are just few observations in the tails. The middle diagram shows that despite these challenges, the conditional survival function is estimated relatively well. The conditional density E-estimate is very good and correctly indicates the underlying conditional density $f^{X^*|X^*>\alpha_{T^*}}(x)$ despite the heavily skewed LTRC observations. The estimated coefficient of difficulty is equal to 2.1, and this, together with the large confidence bands, sheds light on the complexity of the problem and the possibility of poor estimates in other simulations. Nonetheless, it is fair to conclude that the estimators are robust to the above-discussed deviations from the basic LTRC model.

LRTC is one of the most complicated modifications of data, and it is highly recommended to use Figures 6.14 and 6.15 to learn more about this important statistical problem. Exploring different underlying distributions and parameters will help to gain necessary experience in dealing with LTRC data.

6.8 Nonparametric Regression with RC Responses

A classical regression problem, discussed in Section 2.3, is to estimate a regression function

$$m(x) = \mathbb{E}\{Y|X = x\}, \tag{6.8.1}$$

based on a sample of size n from a pair of random variables (X, Y). Here X is the predictor and Y is the response. Recall that $m(x)$ is used to predict the response given $X = x$, and also the regression is a useful tool to describe a relationship between the two variables.

We already know how regression E-estimator is constructed and how it performs for the case of direct observations from (X, Y). In survival analysis it is often the case that one of the variables is modified by censoring. In this section we are considering the case of right censored (RC) responses, and the next one explores the case of right censored predictors.

The following regression model is considered. Response Y is a lifetime (nonnegative random variable) which is right censored by a censoring random variable C. Predictor X is observed directly. As a result, we observe a sample of size n from the triplet (X, V, Δ) where $V := \min(Y, C)$ and $\Delta := I(Y \le C)$. Note that the predictor is not censored and observed directly. In what follows we assume that the pair (X, Y) and the censoring variable C are independent, the three underlying random variables X, Y and C are continuous, and as usual we are assuming that X is supported on $[0, 1]$ according to a continuous and positive design density (the latter yields that $\min_{x\in[0,1]} f^X(x) > 0$ and hence the reciprocal of design density is finite). The problem is to propose, if possible, a consistent estimator of the regression function (6.8.1).

In previous sections a number of classical actuarial, medical and engineering examples of right censoring were presented. Let us add one more example from economics. Consider an observed purchase V which is right censored by rationing C, and Y is the underlying hidden demand that would be equal to the purchase except for the rationing. The rationing means the controlled distribution of scarce resources, goods, or services, or an artificial restriction of demand. We would like to know a relationship between the underlying demand and a predictor of interest, which may be level of inflation, income or unemployment. Note that ignoring censoring would yield a decreased demand, and it is also natural to expect that a severe censoring may preclude us from consistent estimation of the regression.

Now let us present a key probability formula. Following Section 6.5, we can write that

$$f^{X,V,\Delta}(x, y, 1) = f^X(x)f^{Y|X}(y|x)G^C(y). \tag{6.8.2}$$

We also know from Section 6.5 that the distribution of Y cannot be estimated beyond the value $\beta_V = \min(\beta_Y, \beta_C)$ (recall our notation β_Z for the upper bound of the support of a random variable Z). Hence, if

$$\beta_Y < \beta_C, \tag{6.8.3}$$

then the distribution of Y and the regression function (6.8.1) may be consistently estimated, otherwise the distribution of Y may be recovered only up to value $\beta_V = \beta_C$. As a result, let us introduce a censored (or we may say trimmed) regression function

$$m(x, \beta_V) := \mathbb{E}\{YI(Y \leq \beta_V)|X = x\}$$
$$= \int_0^{\beta_V} \frac{yf^{X,Y}(x,y)}{f^X(x)}dy = \int_0^{\beta_V} \frac{yf^{X,V,\Delta}(x,y,1)}{f^X(x)G^C(y)}dy. \tag{6.8.4}$$

In what follows our aim is to estimate the censored regression function (6.8.4) because in general we cannot estimate the regression function (6.8.1).

Note that $m(x, \beta_V) = m(x)$ whenever (6.8.3) holds. In other words, given the assumption (6.8.3), the two characteristics of the relationship between X and Y coincide. But in general $m(x, \beta_V)$ is smaller or equal to $m(x)$, and this is a remark that bears important practical consequences. Indeed, in applications we know neither β_V nor β_Y, and instead are dealing with an empirical $\hat{\beta}_V := V_{(n)}$ which may be significantly smaller than β_Y even if (6.8.3) holds. To understand why note that $G^V(v) = G^Y(v)G^C(v)$ and hence the survival function of V decreases faster than the survival function of Y. As a result, even if assumption (6.8.3) holds, for small samples and depending on severity of censoring, a regression estimate may be significantly smaller than an underlying $m(x)$, and the latter is important to keep in mind when regression with RC response is analyzed. We will complement this discussion shortly by simulated examples.

Now let us explain how to construct E-estimator of the censored regression (6.8.4). Consider the cosine basis $\{\varphi_j(x),\ j = 0, 1, \dots\}$ on $[0, 1]$. The jth Fourier coefficient of the censored regression is

$$\theta_j := \int_0^1 m(x, \beta_V)\varphi_j(x)dx = \int_0^1 \int_0^{\beta_V} \frac{vf^{X,V,\Delta}(x,v,1)\varphi_j(x)}{f^X(x)G^C(v)}dvdx$$
$$= \mathbb{E}\Big\{\frac{\Delta V \varphi_j(X)}{f^X(X)G^C(V)}\Big\}. \tag{6.8.5}$$

The expectation in (6.8.5) implies that we can use a plug-in sample mean estimator. In the expectation, the design density $f^X(x)$ can be estimated by the density E-estimator $\hat{f}^X(x)$ of Section 2.2 (recall that the predictor is observed directly), and the survival function $G^C(v)$ can be estimated by the sample mean estimator (6.5.10),

$$\hat{G}^C(v) := \exp\{-n^{-1}\sum_{l=1}^n (1 - \Delta_l)I(V_l \leq v)/\hat{g}(V_l)\}, \tag{6.8.6}$$

where

$$\hat{g}(v) := n^{-1}\sum_{l=1}^n I(V_l \geq v). \tag{6.8.7}$$

Combining the results, the proposed Fourier estimator is

$$\hat{\theta}_j := n^{-1}\sum_{l=1}^n \frac{\Delta_l V_l \varphi_j(X_l)}{\max(\hat{f}^X(X_l), c/\ln(n))\hat{G}^C(V_l)}. \tag{6.8.8}$$

In its turn, this Fourier estimator allows us to construct the regression E-estimator $\hat{m}(x, \beta_V)$. Further, given (6.8.3) this estimator consistently estimates $m(x)$.

Figure 6.16 allows us to understand the model, observations, and a possible estimation of the regression function. The underlying simulation and diagrams are explained in the caption. The top diagram shows us the available sample from (X, V, Δ) where uncensored cases (when $\Delta = 1$) are shown by circles and censored cases by crosses. The underlying

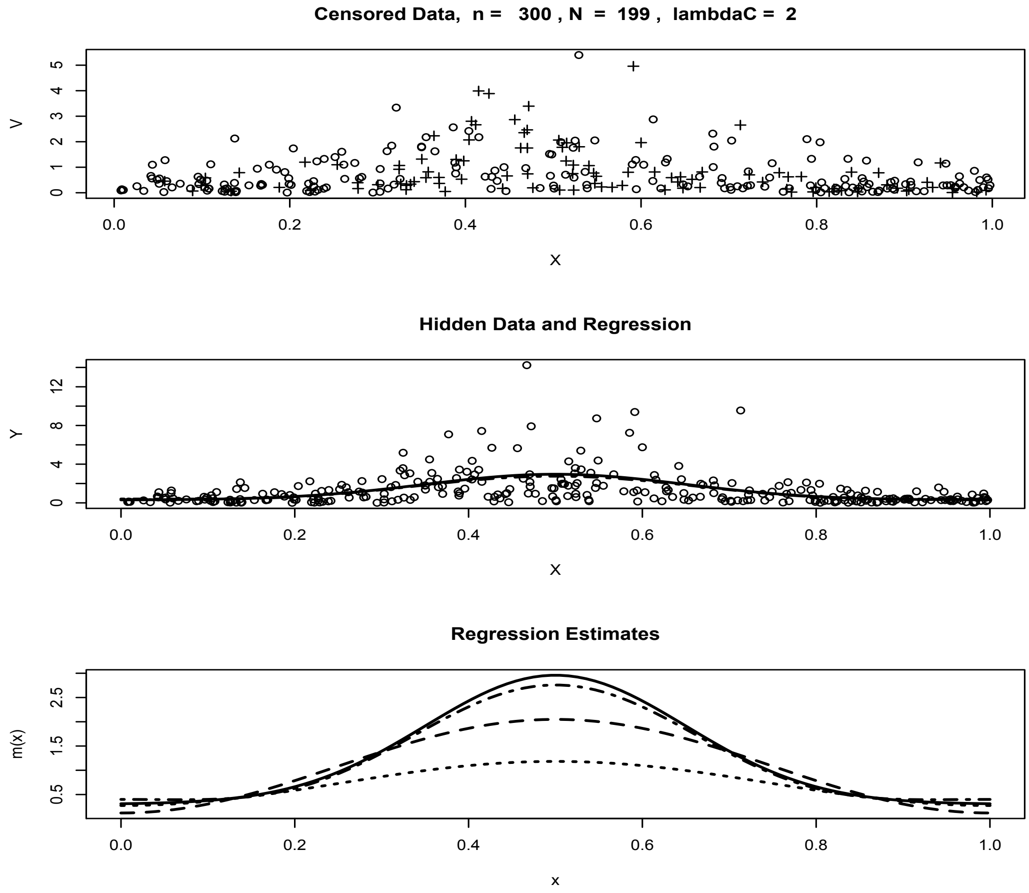


Figure 6.16 *Regression with heavily censored responses. The responses are independent exponential random variables with mean $a + f(X)$ where f is a corner function, here it is the Normal. The predictor is the Uniform. The censoring variable is exponential with the mean λ_C. The top diagram shows by circles observations with uncensored responses and by crosses observations with censored ones, $N := \sum_{l=1}^{n} \Delta_l$. Underlying scattergram of the hidden sample from (X, Y) is shown in the middle diagram, and it is overlaid by the underlying regression function (the solid line) and its E-estimate based on this sample (the dot-dashed line). The bottom diagram shows the underlying regression (the solid line), the E-estimate based on RC data shown in the top diagram (the dashed line), the E-estimate based solely on cases with uncensored responses shown by circles in the top diagram (the dotted line), and the dot-dashed line is the estimate shown in the middle diagram. {Parameter λ_C is controlled by argument lambdaC, function f is chosen by argument corn.} [n = 300, corn = 2, a = 0.3, lambdaC = 2, lambdaC = 2, cJ0 = 4, cJ1 = 0.5, cTH = 4, c = 1.]*

sample from (X, Y) is shown in the middle diagram. It is generated by the Uniform predictor X and Y being exponential variable with the mean equal to $0.3 + f(X)$ where $f(x)$ is the Normal corner function. Note that the mean is the regression function and it is shown by the solid line. Pay attention to the large volatility of the exponential regression and the range of underlying responses. The dot-dashed line is the E-estimate based on the hidden sample, and it may be better visualized in the bottom diagram. The E-estimate is good, and also the scattergram corresponds to a unimodal and symmetric regression function.

Now let us return to the top diagram which shows available data and compare them with the underlying ones shown in the middle diagram. First of all, let us compare the scales.

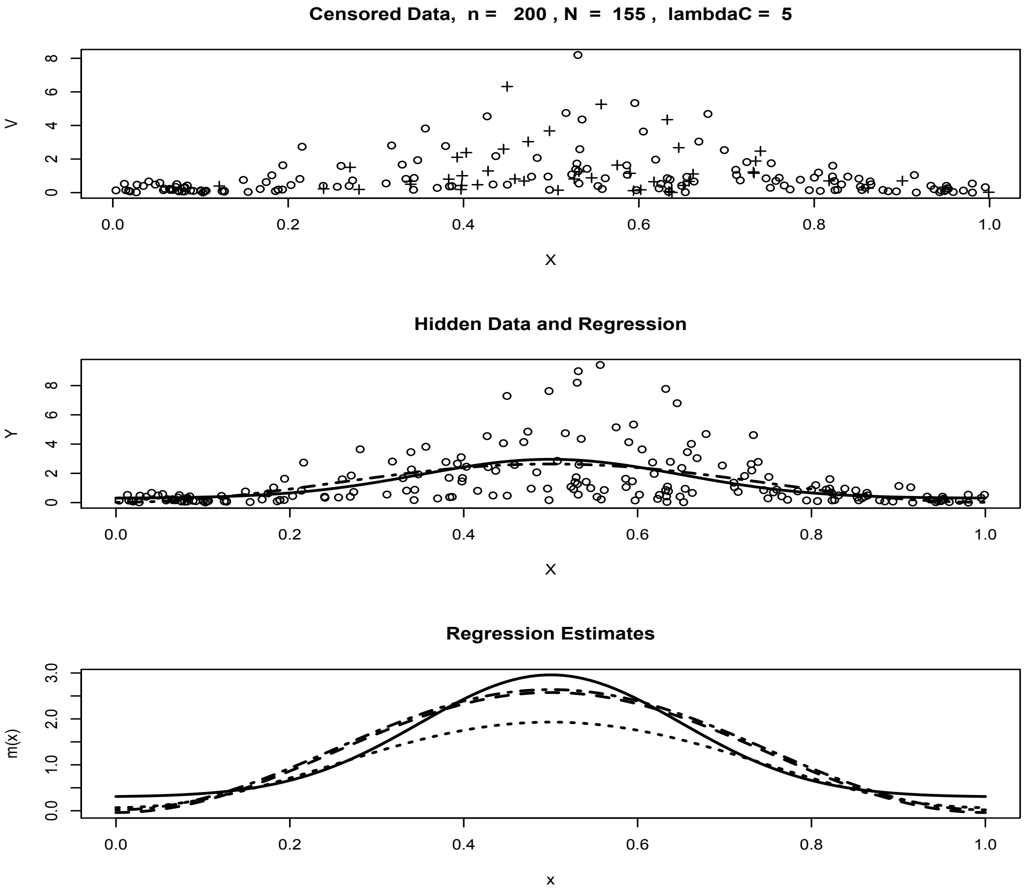


Figure 6.17 *Regression with mildly censored responses. The simulation and diagrams are the same as in Figure 6.16, only here the smaller sample size $n = 200$ and larger mean $\lambda_C = 5$ of the exponential censoring variable are used. [n = 200, corn = 2, a = 0.3, lambdaC = 5, cJ0 = 4, cJ1 = 0.5, cTH = 4, c = 1.]*

Value of the largest observed $V_{(n)}$ is close to 6 while value of the largest $Y_{(n)}$ is close to 14. This is a sign of severe censoring. Further, practically all larger underlying responses are censored, with just few uncensored responses being larger than 3. Further, among $n = 300$ underlying responses, only $N = 199$ are uncensored, that is every third response is censored. If we return to our discussion of the assumption (6.8.3) and the censored regression, we may expect here that a regression estimate, based on the censored data, may be significantly smaller than the underlying regression. The bottom diagram supports this prediction. The dashed line is the E-estimate, and it indeed significantly underestimates the underlying regression function shown by the solid line. Nonetheless, the proposed estimator does a better job than a complete-case approach which yields an estimate shown by the dotted line. Recall that a complete-case approach simply ignores cases with censored responses. What we see is that despite a relatively large sample size, the observed $V_{(n)}$ is too small and this explains the underperformance of the E-estimator.

Of course, the simulation of Figure 6.16 implies a severe censoring. Let us relax it a bit by considering a censoring variable with exponential distribution only now with a larger mean $\lambda_C = 5$. This should produce more uncensored observations and a better regression estimation. A particular outcome is shown in Figure 6.17 where also a reduced sample size $n = 200$ is chosen for better visualization of scattergrams. Otherwise, the simulation and diagrams are the same as in Figure 6.16.

Let us begin with comparison of observations in the top and middle diagrams. First of all, note that the scales are about the same, and $V_{(n)}$ is close to $Y_{(n)}$. Further, now less than a quarter of responses is censored. This is still a significant proportion but dramatically smaller than in Figure 6.16. Now let us look at the estimates. The E-estimates based on the underlying and censored data are close to each other (compare dot-dashed and dashed curves), and they are much better than the E-estimate based on complete cases (the dotted line).

It is highly recommended to repeat these two figures with different parameters and gain experience in dealing with censored responses. It is useful to manually analyze scattergrams and try to draw a reasonable regression which takes into account the nature of data. Further, use different regression functions and realize whether the E-estimator allows us to make correct conclusions about modes, namely about their number, locations and relative magnitudes.

6.9 Nonparametric Regression with RC Predictors

Consider a regression problem where the predictor X is right censored by a censoring variable C and the response Y is observed directly. In this case we observe a sample from the triplet (U, Y, Δ) where $U := \min(X, C)$ and $\Delta = I(X \le C)$. The aim is to estimate an underlying regression

$$m(x) := \mathbb{E}\{Y|X = x\}. \tag{6.9.1}$$

This is an interesting and possibly complicated setting because we are dealing with a modified predictor, and we know from Chapter 4 that typically this implies serious complications. And indeed, as we will see shortly, right censored predictors may preclude us from consistent estimation. At the same time, if possible a consistent estimation is surprisingly simple.

We begin our analysis of the problem with probability formulas. Assume that the pair (X, Y) and C are independent, the three variables are continuous and nonnegative, and as usual the predictor X has a continuous and positive density on the support $[0, 1]$.

Then we can write that

$$f^{U,Y,\Delta}(x, y, 1) = f^X(x) f^{Y|X}(y|x) G^C(x). \tag{6.9.2}$$

This formula allows us to get the joint density of (U, Y) given $\Delta = 1$, that is the joint density of observations for uncensored (complete) cases. Write,

$$f^{U,Y|\Delta}(x, y|1) = \frac{f^X(x) f^{Y|X}(y|x) G^C(x)}{\mathbb{P}(\Delta = 1)}$$

$$= \Big[\frac{f^X(x)G^C(x)}{\int_0^1 f^X(u)G^C(u)du}\Big] f^{Y|X}(y|x) =: f^Z(x) f^{Y|X}(y|x). \tag{6.9.3}$$

Here $f^Z(x)$ is the density of uncensored predictors, that is $f^Z(x) := f^{U|\Delta}(x|1)$.

Relation (6.9.3) is the key for understanding the data and proposed regression E-estimator. The relation tells us that, whenever $f^Z(x)$ is positive, the conditional density of Y given $U = x$ in an uncensored case, that is given $\Delta = 1$, is the same as the underlying conditional density of Y given $X = x$. As a result, for those x the uncensored-case approach is consistent. Further, even estimation of the conditional density may be based on that approach. On the other hand, if $\beta_C < \beta_X$, then $f^Z(x) = 0$ for $x \in [\beta_C, \beta_X]$ and we cannot consistently estimate a regression function over that interval. The latter is the

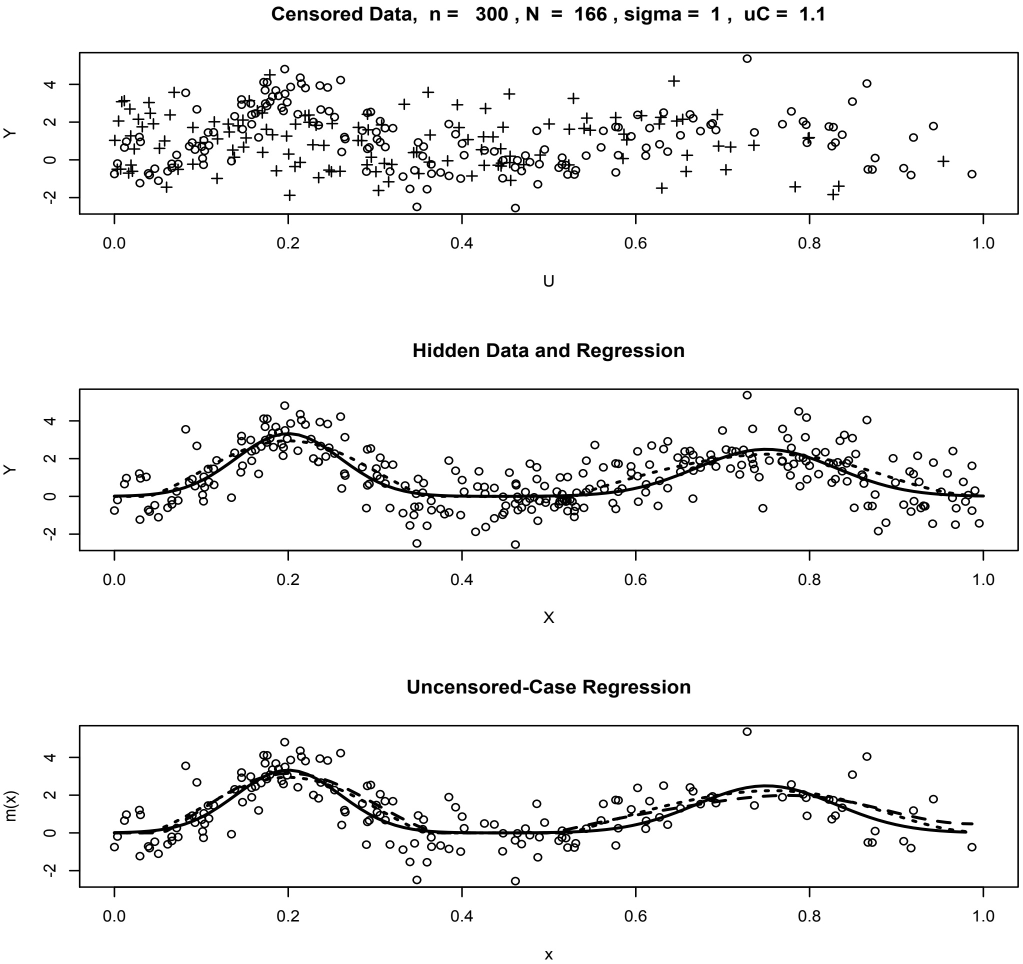


Figure 6.18 *Regression with censored predictors. The underlying predictors have the Uniform distribution, the responses are $m(X)+\sigma\varepsilon$ where $m(x)$ is the Strata and ε is standard normal, and the censoring variable is Uniform$(0, u_C)$. The top diagram shows by circles observations with uncensored predictors and by crosses observations with censored predictors, $N := \sum_{l=1}^{n} \Delta_l$ is the number of uncensored predictors. Underlying scattergram of the hidden sample from (X, Y) is shown in the middle diagram, and it is overlaid by the underlying regression function (the solid line) and its E-estimate based on this sample (the dotted line). The bottom diagram shows data and curves over an interval $[0, U_{(n)}]$. Circles show observations with uncensored predictors (they are identical to circles in the top diagram). These observations are overlaid by the underlying regression (the solid line), the E-estimate based on uncensored-case observations (the dashed line), and the E-estimate based on underlying observations (the dotted line) and it is the same as in the middle diagram. {The regression function is chosen by argument corn, parameter u_C is controlled by the argument uC.} [n = 300, corn = 4, sigma = 1, uC = 1.1, cJ0 = 4, cJ1 = 0.5, cTH = 4]*

familiar curse of severe censoring. On the other hand, we may always consistently estimate left tail of regression.

Let us compare our conclusions for regressions with MAR data (recall that the latter was discussed in Chapter 4). For the setting of MAR responses a complete-case approach is optimal, and for the setting of MAR predictors a special estimator, which uses all obser-

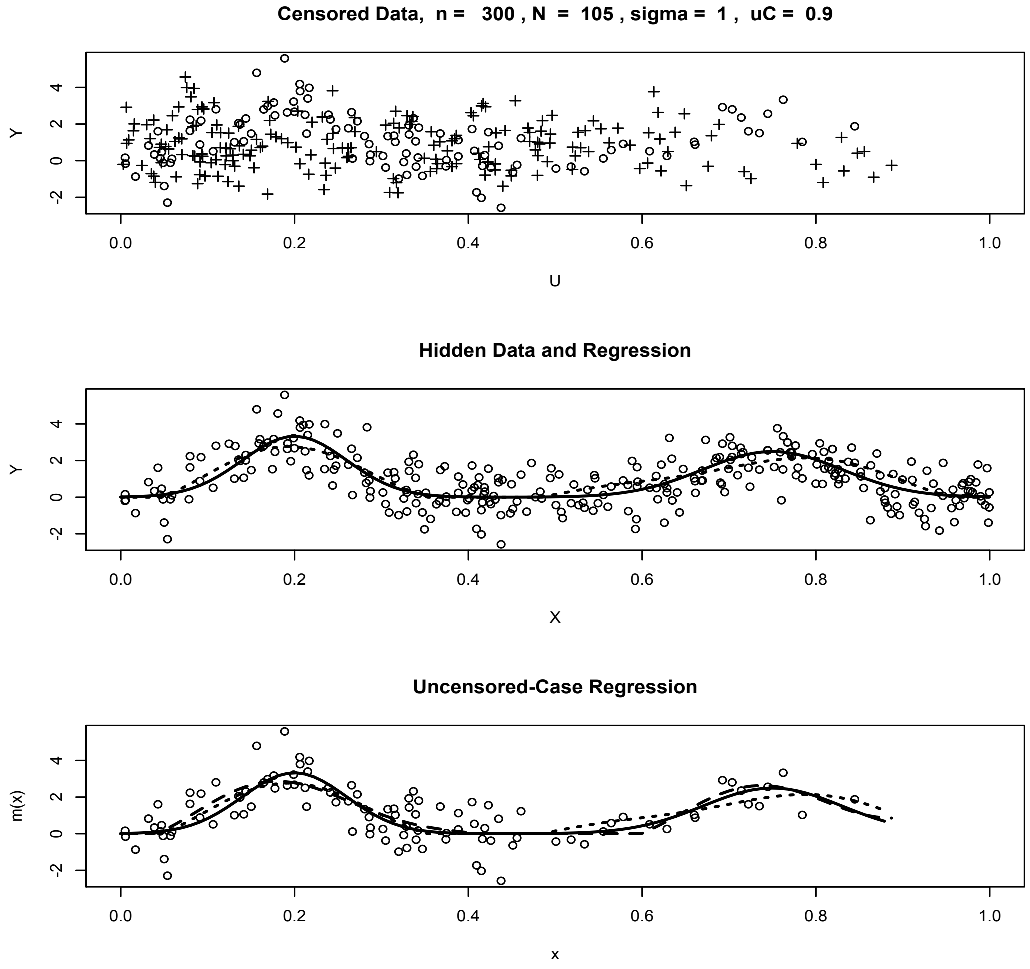


Figure 6.19 *Regression with heavily censored predictors. The simulation and the diagrams are the same as in Figure 6.18 only here $u_C = 0.9$. [n = 300, corn = 4, sigma = 1, uC = 0.9, cJ0 = 4, cJ1 = 0.5, cTH = 4]*

vations, is needed. As we now know, for regression with RC data the outcome is different. Here a complete-case approach is optimal for the setting of censored predictors and a special estimator, based on all observations, is needed for the setting of censored responses. This is a teachable moment because each modification of data has its own specifics and requires a careful statistical analysis.

Figure 6.18 allows us to understand the discussed regression with right censored predictor and the proposed solution. Its caption explains the simulation and the diagrams. The top diagram shows the scattergram of censored observations with circles indicating cases with uncensored predictors and crosses indicating cases with censored predictors. Note that the largest available predictor is near 1, and this hints that the censoring variable may take even larger values. And indeed, the censoring variable is supported on $[0, u_C]$ with $u_C = 1.1$. Further, while the underlying predictor X has the Uniform distribution, the observed U is left skewed. We see this in the diagram, and this also follows from the relation $G^U(x) = G^X(x)G^C(x)$. The latter allows us to conclude that RC predictors may create problems for estimation of a right tail of the regression. Further, note that 45% of predictors are censored, and this is a significant loss of data.

The middle diagram shows us the underlying (hidden) scattergram, the regression function and its E-estimate. The estimate is not perfect, but note the large standard deviation σ=1 of the regression error (controlled by the argument sigma) and recall that the Strata is a difficult function for a perfect estimation. The bottom diagram sheds light on the proposed uncensored-case approach. If we look at the exhibited uncensored cases, shown by circles, then we may visualize the underlying Strata regression. Of course, just few observations in the right tail complicate the visualization, but the E-estimator does a good job and the E-estimate (the dashed line) is comparable with the estimate based on all hidden observations (the dotted line).

What will be if we use a more severe censoring, for instance one with $u_C = 0.9$? A corresponding outcome is shown in Figure 6.19. The top diagram shows that the largest uncensored predictor is about 0.85, while the largest observation of the censoring variable is near 0.9. This tells us that it is likely that the support of U is defined by the censoring variable C, and this is indeed the case here. Further, note that we have just few uncensored predictors with values larger than 0.75. Further, note that only $N = 105$ predictors from underlying $n = 300$ are uncensored and they are heavily left skewed. What we see is an example of a severe censoring.

The middle diagram shows that the sample of underlying hidden observations is reasonable and, keeping in mind the large regression noise (it is possible to reduce or increase it using the argument sigma), the E-estimate is fair. The bottom diagram shows us the E-estimate based on uncensored cases (the dashed line) which can be compared with the underlying regression (the solid line) and the E-estimate of the middle diagram (the dotted line). As we see, despite the small sample size $N = 105$ of uncensored cases and the large regression noise, the uncensored-case E-estimate is relatively good. Let us stress that it is calculated for the interval $[0, U_{(n)}]$, and that we have no information about the underlying regression beyond this interval.

Repeated simulations of Figure 6.18, using different parameters and underlying regressions, may help to shed a new light on the interesting and important statistical problem of regression with censored predictor.

6.10 Exercises

6.1.1 Verify (6.1.4).
6.1.2 Prove that if the hazard rate is known, then the survival function can be calculated according to formula (6.1.3).
6.1.3 Verify (6.1.5).
6.1.4 Consider the Weibull distribution with the shape parameter k defined below line (6.1.7). Prove that if $k < 1$ then the hazard rate is decreasing and if $k > 1$ then it is increasing.
6.1.5* Suppose that X and Y are two independent lifetimes with known hazard rates and $Z := \min(X, Y)$. Find the hazard rate of Z.
6.1.6 Propose a hazard rate whose shape resembles a bathtub.
6.1.7* Prove that for any hazard rate h^Y the relation $\int_0^\infty h^Y(y)dy = \infty$ holds. Furthermore, if S is the support of Y then $\int_S h^Y(y)dy = \infty$.
6.1.8 Prove that $G^Y(y) = G^Y(a)e^{-\int_a^y h^Y(u)du}$ for any $a \in [0, y]$.
6.1.9 Verify (6.1.6) and (6.1.7). Hint: Here the scale-location transformation is considered. Begin with the cumulative distribution function and then take the derivative to get the corresponding density.

6.1.10* Propose a series estimator of the hazard rate which uses the idea of transformed $Z = (X - a)/b$.

6.1.11 Verify expression (6.1.8) for Fourier coefficients of the hazard rate.

6.1.12 Prove that the oracle-estimator $\tilde{\theta}_j^*$, defined in (6.1.10), is unbiased estimator of θ_j.

6.1.13 Find variance of the estimator (6.1.10). Hint: Note that the oracle is a classical sample mean estimate, and then use formula for the variance of the sum of independent random variables.

6.1.14 Find the mean and variance of the empirical survival function (6.1.14). Hint: Note that this estimate is the sample mean of independent and identically distributed Bernoulli random variables. Furthermore, the sum has a Binomial distribution.

6.1.15* Use Hoeffding's inequality for the analysis of large deviations of the empirical survival function (6.1.14).

6.1.16* Evaluate the mean and variance of the Fourier estimator (6.1.15). Hint: Begin with the case of known nuisance functions, consider the asymptotic as n and j increase, and then show that the plug-in methodology is valid.

6.1.17* Explain why (6.1.17) is a reasonable estimator of the coefficient of difficulty. Hint: Replace the estimated survival function by an underlying survival function G^Y, and then calculate the mean and variance of this oracle-estimator.

6.1.18* Explain why the ratio-estimator (6.1.19) may be a good estimator of the hazard rate. What are possible drawbacks of the estimator?

6.1.19 Repeat Figure 6.1 for sample sizes n =100, 200, 300, 400, 500 and make your conclusion about corresponding feasible intervals of estimation. Comment about your method of choosing an interval. Does a sample size affect the choice of interval?

6.1.20 Repeat Figure 6.1 twenty times, write down ISEs for the estimates, and then rank the estimates based on results of this numerical study.

6.1.21 Repeat Figure 6.1 for different underlying distributions. Choose largest feasible intervals of estimation for each distribution.

6.1.22 What are the optimal parameters of the E-estimator for the experiment considered in Figure 6.1?

6.1.23 Explain underlying simulations and all histograms in Figure 6.1.

6.1.24* Write down the hazard rate E-estimator and explain parameters and statistics used.

6.2.1 Give several examples of right censoring. Further, present an example of left censoring and explain the difference.

6.2.2 Explain the mechanism of right censoring. How are available observations in right-censored data related to underlying random variables? Hint: Check (6.2.1).

6.2.3 Are random variables V and Δ dependent, independent or is there not enough information to answer?

6.2.4 Explain formula (6.2.2). Why is its last equality valid?

6.2.5 Prove (6.2.3). Formulate assumptions needed for its validity.

6.2.6 Formulate assumptions and verify (6.2.4). Describe the support of the density. Is this a mixed density? What is the definition of a mixed density?

6.2.7 Are censored observations biased? If the answer is "yes," then what is the biasing function?

6.2.8 Explain formula (6.2.5). Does it express the hazard rate via functions that can be estimated?

6.2.9 What is the motivation behind expression (6.2.6) for Fourier coefficients?

6.2.10* Explain the Fourier estimator (6.2.7) and then describe its statistical properties. Hint: Find the mean, the variance and the probability of large deviations using the Hoeffding inequality.

6.2.11* What is the coefficient of difficulty? What is its role in nonparametric estimation?

6.2.12* Verify (6.2.9) and (6.2.10).

6.2.13 Explain the simulation used to create Figure 6.2.
6.2.14 Repeat Figure 6.2 and write down analysis of its diagrams.
6.2.15 Describe all parameters/arguments of Figure 6.2. What do they affect?
6.2.16 Repeat Figure 6.2 about 20 times and make your own conclusion about the effect of estimator $\hat{G}^V$ on the hazard rate E-estimator.
6.2.17 Explain how the interval of estimation affects the hazard estimator. Support your conclusion using Figure 6.2.
6.2.18 Consider different distributions of the censoring variable. Then, using Figure 6.2, present a report on how those distributions affect a reasonable interval of estimation and the quality of E-estimator.
6.2.19* Explain formula (6.2.11) for the empirical coefficient of difficulty. Then calculate its mean and variance.
6.2.20* Propose a hazard rate E-estimator for left censored data.

6.3.1 Give an example of left truncated data.
6.3.2 Describe an underlying stochastic mechanism of left truncation. Is it a missing mechanism?
6.3.3 Formulate the probabilistic model of truncation.
6.3.4 Assume that the sample size of a hidden sample is n. Is the sample size of a corresponding truncated sample deterministic or stochastic? Explain your answer.
6.3.5 What are the mean and variance of the sample size of a truncated sample given that n is the size of an underlying hidden sample?
6.3.6 Explain and verify (6.3.3). What are the used assumptions?
6.3.7 Explain and verify (6.3.4). Hint: Pay attention to the support of this bivariate density.
6.3.8* Assume that T is a discrete random variable. How will (6.3.3) and (6.3.4) change?
6.3.9 Verify (6.3.5).
6.3.10* Explain and verify (6.3.6). Pay attention to the conditions when the equality holds. Will the equality be valid for points $x \leq \alpha_{T^*}$?
6.3.11 Does truncation imply biasing? If the answer is "yes," then what is the biasing function?
6.3.12 Does left truncation skew a hidden sample of interest to the left or the right?
6.3.13* Explain the motivation of introducing the probability $g(x)$ in (6.3.8). Hint: Can this function be estimated based on a truncated sample? Then look at (6.3.9).
6.3.14* Verify relations in (6.3.8).
6.3.15 Verify (6.3.9).
6.3.16* How can (6.3.9) be used for estimation of the hazard rate of interest?
6.3.17 Explain the underlying idea of Fourier estimator (6.3.10). Hint: Replace the estimate $\hat{g}$ by g and show that this is a sample mean estimator.
6.3.18* What is the mean of the estimator (6.3.10)? Is it unbiased or asymptotically unbiased? Calculate the variance.
6.3.19 Explain the motivation behind the estimator (6.3.11).
6.3.20 What is the mean and variance of the estimator (6.3.11)?
6.3.21* Use Hoeffding's inequality for the analysis of the estimator (6.3.11). Then explain why this estimator may be used in the denominator of (6.3.10).
6.3.22* Verify (6.3.12).
6.3.23 Explain the motivation behind the estimator (6.3.13).
6.3.24* What are the mean and the variance of the estimator (6.3.13).
6.3.25* Explain how an E-estimator of the hazard rate is constructed.
6.3.26 Describe the underlying simulation that creates data in Figure 6.3.
6.3.27 Repeat Figure 6.3 a number of times and recommend a feasible interval of estimation.
6.3.28 Using Figure 6.3, propose "good" values for parameters of the E-estimator.

6.3.29 Using Figure 6.3, explore the issue of how truncating variables affect quality of estimation of an underlying hazard function.
6.3.30 Can Figure 6.3 be used for statistical analysis of the used confidence bands? Test your suggestion.
6.3.31 Rank corner distributions according to difficulty in estimation of their hazard rates. Then check your conclusion using Figure 6.3.
6.3.32 Confidence bands may take on negative values. How can they be modified to take into account that a hazard rate is nonnegative?
6.3.33* Consider three presented examples of left censoring (actuarial, startups and clinical trials), and explain how each may be "translated" into another.
6.3.34* Write down the hazard rate E-estimator and explain parameters and statistics used.

6.4.1 Give several examples of LTRC data.
6.4.2 Explain an underlying stochastic mechanism of creating LTRC data.
6.4.3 Present examples when the LT is followed by the RC and vice versa.
6.4.4 Explain formula (6.4.1). Can the probability be estimated?
6.4.5 Suppose that the sample size of an underlying hidden sample of interest is n. What is the distribution of the sample size of a LTRC sample? What are its mean and variance?
6.4.6 Present examples when truncating and censoring variables are dependent and independent.
6.4.7* Explain formula (6.4.2) for the cumulative distribution function of the triplet of random variables observed in a LTRC sample. Comment on the support. What does the formula tell us about a possibility of consistent estimation of the distributions of X^*, T^* and C^*?
6.4.8 Verify formula (6.4.3) and explain assumptions under which it is correct.
6.4.9 Prove validity of (6.4.4).
6.4.10 What is the meaning of the joint density (6.4.4)? Hint: note that one of the variables is discrete.
6.4.11 Verify expression (6.4.5) for the marginal mixed density of (V, Δ).
6.4.12* Verify validity of formula (6.4.6) for the density of the variable of interest X^*. Explain the assumptions when it is valid. Can this formula be used for construction of a consistent E-estimator and what are the assumptions?
6.4.13* Why is probability (6.4.7) a pivotal step in constructing an E-estimator?
6.4.14 Explain formula (6.4.7).
6.4.15 Can the function $g(x)$ be estimated based on LTRC data?
6.4.16 Verify formulas (6.4.8) and (6.4.9). Explain the assumptions.
6.4.17* Suggest an estimator for $h^{X^*}(x)$. Hint: Use (6.4.9).
6.4.18 Is the Fourier estimator (6.4.10) a sample mean estimator?
6.4.19 Why can $\hat{g}(x)$, defined in (6.4.11), be used in the denominator of (6.4.10)?
6.4.20* Present a theoretical statistical analysis of estimators (6.4.10) and (6.4.11). Hint: Begin with the distribution and then write down the mean, the variance, and for estimator (6.4.11) use the Hoeffding inequality to describe large deviations.
6.4.21 Verify (6.4.12).
6.4.22 Explain the estimator (6.4.13). Is it asymptotically unbiased?
6.4.23 Conduct a series of simulations, using Figure 6.4, and explore the effect of estimate $\hat{g}$ on the E-estimate of the hazard rate.
6.4.24 Choose a set of sample sizes, truncated and censoring distributions, and then try to determine a feasible interval of estimation for each underlying model. Explain your choice.
6.4.25 Suggest better values of parameters of the E-estimator. Does your recommendation depend on distributions of truncating and censoring variables?
6.4.26 What is the main difference between simulations in Figures 6.4 and 6.5?
6.4.27 Present several examples where T^* and C^* are related.

6.4.28 Present several examples where $\mathbb{P}(C^* \geq T^*) = 1$ and C^* has a mixed distribution.
6.4.29* Verify each equality in (6.4.14). Explain the used assumptions.
6.4.30 Establish validity of (6.4.16). What are the used assumptions?
6.4.31 Prove (6.4.17).
6.4.32 Explain (6.4.18) and the underlying assumptions.
6.4.33 Describe the underlying simulation of Figure 6.5.
6.4.34 Explain diagrams in Figure 6.5.
6.4.35 For estimation of the hazard rate, is the model of Figure 6.5 more challenging than of Figure 6.4?
6.4.36 Using Figure 6.5, suggest better parameters of the E-estimator.
6.4.37 Using Figure 6.5, infer about performance of the confidence bands.
6.4.38* Consider a setting where the made assumptions are violated. Then propose a consistent hazard rate estimator or explain why this is impossible.

6.5.1 Present several examples of RC observations.
6.5.2 Is there something in common between RC and MNAR? If the answer is "yes," then why does MNAR typically preclude us from a consistent estimation and RC not?
6.5.3 Explain validity of (6.5.1) and the underlying assumption.
6.5.4 Prove (6.5.2).
6.5.5 Under RC, over what interval may the distribution of interest be estimated?
6.5.6 Describe assumptions that are sufficient for consistent estimation of the distribution of X.
6.5.7 Explain how the Kaplan–Meier estimator is constructed.
6.5.8* What is the underlying motivation of the Kaplan–Meier estimator? Why is it called a product limit estimator? Evaluate its mean and variance.
6.5.9 Explain formulae (6.5.6) and (6.5.7). Under what assumptions are they valid?
6.5.10 Explain how to estimate the survival function of the censoring random variable C.
6.5.11 Verify (6.5.9).
6.5.12* Find the mean and variance of the estimator (6.5.10).
6.5.13 Explain and then verify (6.5.12).
6.5.14 What is the motivation behind the Fourier estimator (6.5.13)?
6.5.15* Find the mean and variance of the estimator (6.5.13).
6.5.16 Using (6.5.14) explain how the estimator (6.5.15) is constructed.
6.5.17 Explain the simulation of Figure 6.6.
6.5.18 Explain the simulation of Figure 6.7.
6.5.19 What is the difference between simulations in Figures 6.6 and 6.7?
6.5.20 Propose better values of parameters for E-estimators used in Figures 6.6 and 6.7. Are they different? Explain your findings.
6.5.21* Consider a setting where some of the made assumptions are no longer valid. Then explore a possibility of consistent estimation.

6.6.1 Describe the model of LT.
6.6.2 Present several examples of LT observations. Based solely on LT observations, can one conclude that the observations are LT?
6.6.3 Is LT based on a missing? Is the missing MNAR? Typically MNAR precludes us from consistent estimation. Is the latter also the case for LT?
6.6.4 Explain the assumption (6.6.1) and its importance.
6.6.5 Why is the assumption (6.6.2) important?
6.6.6 Verify each equality in (6.6.3) and explain used assumptions.
6.6.7 Verify (6.6.5) and explain the used assumption.
6.6.8 Establish (6.6.6) and (6.6.7).
6.6.9* Explain the underlying idea of estimators defined in (6.6.9). Find their expectations. Can these estimators be improved?

6.6.10 Verify all relations in (6.6.10).
6.6.11* Explain construction of the estimator (6.6.11). Find its mean and variance.
6.6.12* Conduct a statistical analysis of the estimator (6.6.12). Hint: Describe the distribution and its properties.
6.6.13* Explain the motivation behind the estimator (6.6.14). Evaluate its mean and variance.
6.6.14 Verify (6.6.16).
6.6.15* Suggest an E-estimator of the density f^{X^*}.
6.6.16* Find the mean and variance of Fourier estimator (6.6.19).
6.6.17 What is the motivation behind the estimator (6.6.21)? Can you propose another feasible estimator?
6.6.18 Explain the simulation used by Figure 6.8.
6.6.19 Repeat Figure 6.8 and analyze diagrams.
6.6.20 Use Figure 6.8 and compare performance of the Kaplan–Meier estimator with the sample mean estimator.
6.6.21 Explain the underlying idea of estimator (6.6.22).
6.6.22 Use Figure 6.9 to compare performance of the two estimators.
6.6.23 Explain all relations in (6.6.23).
6.6.24 Explain diagrams in Figure 6.10. Then use it for statistical analysis of the proposed density estimator.
6.6.25 Explain the underlying simulation in Figure 6.11.
6.6.26 Repeat Figure 6.11 for different sample sizes. Write a report about your findings.
6.6.27 Suggest better values for parameters of the E-estimator used in Figure 6.11. Is your recommendation robust to changes in other arguments of the figure?
6.6.28 Explain how the cumulative distribution function of T^* can be estimated.
6.6.29 Explain how the probability density of T^* can be estimated.
6.6.30 Describe the underlying simulation in Figure 6.12.
6.6.31 Explain diagrams in Figure 6.12.
6.6.32 Describe E-estimators used in Figure 6.12.
6.6.33 Explain formula (6.6.29).
6.6.34* Evaluate the mean and variance of Fourier estimator (6.6.30).
6.6.35 Prove (6.6.31).
6.6.36 Verify (6.6.32), and then explain why we are interested in the analysis of $\mathbb{E}\{\hat{H}^{X^*}(x)\}$.
6.6.37* Show that the second expectation in the right side of (6.6.32) vanishes as n increases. Hint: Propose any needed assumptions.
6.6.38 Verify (6.6.33).
6.6.39 Prove (6.6.34). Then explain meaning of the conditional survival function.
6.6.40 Given $\alpha_{X^*} < \alpha_T^*$, explain why the distribution of X^* cannot be consistently estimated, and then explain what may be estimated.
6.6.41 Explain all curves in the middle diagram of Figure 6.11. Then repeat it and analyze the results. Is estimation of the conditional survival function robust?
6.6.42 What does $\hat{p}$ estimate in Figure 6.11?
6.6.43 Explain each relation in (6.6.35).
6.6.44* Show that the second expectation on the right side of (6.6.35) vanishes as n increases. Hint: Make a reasonable assumption.
6.6.45 Verify (6.6.36).
6.6.46 Explain the definition of conditional density (6.6.37). Then comment on what can and cannot be estimated given $\alpha_{T^*} > \alpha_{X^*}$.
6.6.47 Consider the bottom diagram in Figure 6.11 and explain the curves. Then repeat Figure 6.11 with different parameters and explore robustness of the proposed estimators.
6.6.48* Explain all steps in construction of the E-estimator $\hat{f}^{T^*}(t)$. Formulate necessary assumptions.

6.6.49 Use Figure 6.13 and analyze statistical properties of the confidence bands.
6.6.50 Is the relation $\mathbb{P}(T_{(n)} \leq X_{(n)}) = 1$ valid?
6.6.51 Is there any relationship between $X_{(1)}$ and $T_{(1)}$?
6.6.52* Relax one of the used assumptions and propose a consistent estimator of $\hat{f}^{X^*}(x)$ or prove that the latter is impossible.

6.7.1 Present examples of left truncated, right censored and LTRC data.
6.7.2 What is the difference (if any) between truncated and censored data.
6.7.3 Explain how LTRC data may be generated.
6.7.4 Find a formula for the probability of an observation in a LTRC simulation.
6.7.5 Explain each equality in (6.7.2).
6.7.6* Using (6.7.3), obtain formulas for corresponding marginal densities.
6.7.7 Can the probability (6.7.7) be estimated based on LTRC data?
6.7.8* Explain assumptions (6.7.8). What will be if they do not hold?
6.7.9 Why is formula (6.7.9) critical for suggesting a density E-estimator?
6.7.10 Explain how formula (6.7.10) is obtained.
6.7.11* What is the motivation behind the estimator (6.7.11)? Find its mean and variance.
6.7.12* Present statistical analysis of the estimator $\hat{g}$. Hint: Think about its distribution.
6.7.13 Verify (6.7.13).
6.7.14 Explain all relations in (6.7.14).
6.7.15* Is (6.7.15) a sample mean Fourier estimator? Find its mean and variance.
6.7.16* Verify (6.7.16).
6.7.17 Explain how underlying distributions of the truncating and censoring random variables affect the coefficient of difficulty of the density E-estimator.
6.7.18 Explain the underlying simulation used in Figure 6.14.
6.7.19 Explain diagrams in Figure 6.14.
6.7.20 Using Figure 6.14, present statistical analysis of the E-estimator.
6.7.21 How well do the confidence bands perform? Hint: Use repeated simulations of Figure 6.14.
6.7.22 Explain every argument of Figure 6.14.
6.7.23* Write down a report about the effect of distributions of the truncated and censoring variable on quality of E-estimate. Hint: Begin with the theory based on the coefficient of difficulty and then complement your conclusion by empirical evidence created with the help of Figure 6.14.
6.7.24 What parameters of the E-estimator, used in Figure 6.14, would you recommend for sample sizes $n = 100$ and $n = 300$?
6.7.25 Use Figure 6.15 and explain how the dependence between truncated and censored variables affects the estimation.
6.7.26 Explain the motivation behind the model (6.7.17). Present several corresponding examples.
6.7.27 Explain and verify each equality in (6.7.18).
6.7.28 Verify every equality in (6.7.19).
6.7.29 Explain how formula (6.7.21) for the density of interest is obtained.
6.7.30* Using (6.7.21), suggest an E-estimator of the density. Hint: Describe all steps and assumptions.
6.7.31* Consider the case $\alpha_{X^*} < \alpha_{T^*}$ and develop the theory of estimation of the conditional survival function $G^{X^*|X^*>\alpha_{T^*}}(x)$.
6.7.32* Consider the case $\alpha_{X^*} < \alpha_{T^*}$ and develop the theory of estimation of the conditional density $f^{X^*|X^*>\alpha_{T^*}}(x)$.
6.7.33 Using Figure 6.14, explore the proposed E-estimators for the case $\alpha_{X^*} < \alpha_{T^*}$.

6.8.1 Present an example of a regression problem with direct observations. Then describe a situation when response may be censored.
6.8.2 Explain (6.8.2).
6.8.3* What is the implication, if any, of assumption (6.8.3)? What can be done, if any, if (6.8.3) does not hold?
6.8.4 What is the meaning of the censored regression function (6.8.4)? Verify each equality in (6.8.4).
6.8.5* Explain the underlying idea of consistent estimation of the regression.
6.8.6 Verify relations in (6.8.5).
6.8.7* Explain the estimator (6.8.6). Evaluate its mean and variance.
6.8.8 What is the distribution of estimator (6.8.7)?
6.8.9 Consider a RC data with Y being the random variable of interest and C being the censoring random variable. If $\hat{G}^Y(y)$ is an estimator of the survival function of Y, how can this estimator be used for estimation of $G^C(z)$?
6.8.10 Explain the estimator (6.8.8).
6.8.11* What is the mean and variance of the Fourier estimator (6.8.8)?
6.8.12* Consider a setting where the made assumptions do not hold. Then explore a possibility of consistent regression estimation.
6.8.13 Consider diagrams in Figure 6.16 and explain the underlying simulation.
6.8.14 What are the four curves in the bottom diagram of Figure 6.16? Why are they all below the underlying regression? Is this always the case?
6.8.15* Explain theoretically how the parameter λ_C affects the regression estimation, and then compare your conclusion with empirical results using Figure 6.16.
6.8.16 Suggest better parameters of the E-estimator for Figure 6.16.
6.8.17 Conduct several simulations similar to Figures 6.16 and 6.17, and then explain the results.
6.8.18 Do you believe that values of parameters of the E-estimator should be different for simulations shown in Figures 6.16 and 6.17? If the answer is "yes," then develop a general recommendation for choosing better values of the parameters.

6.9.1 Explain the mechanism of RC modification. Does this modification involve a missing mechanism? If "yes," then is it MAR or MNAR?
6.9.2 Present several examples of a regression with RC predictor.
6.9.3 What complications in regression estimation may be expected from RC predictor?
6.9.4* Write down probability formulas for all random variables involved in regression with RC predictor.
6.9.5 Can RC predictor imply a destructive modification when a consistent regression estimation is impossible?
6.9.6 For the case of RC response, the notion of a censored regression was introduced. Is there a need to use this notion for the case of RC predictor?
6.9.7 Explain a difference (if any) between regressions with censored predictor and response.
6.9.8 Is expression (6.9.2) correct? Do you need any assumptions? Prove your assertion.
6.9.9 Verify every equality in (6.9.3). Do you need any assumptions for its validity?
6.9.10* Explain why the relation (6.9.3) is the key in regression estimation.
6.9.11* Describe the random variable Z defined in (6.9.3). Propose an estimator of its density.
6.9.12* Propose an E-estimator of the conditional density $f^{Y|X}(y|x)$.
6.9.13 What are the assumptions for consistent estimation of the regression?
6.9.14 Explain the underlying simulation used in Figure 6.18.
6.9.15* Explain, step by step, how the regression E-estimator, used in Figure 6.18, is constructed.

6.9.16* Explore theoretically and empirically, using Figure 6.18, the effect of parameter u_C on estimation.
6.9.17 In your opinion, which of the corner functions are less and more difficult for estimation? Hint: Use Figure 6.18.
6.9.18 Repeat Figure 6.19. Comment on scattergrams and estimates.
6.9.19 Propose better values for parameters of the estimators used in Figures 6.18 and 6.19. Explain your recommendation. Is it robust toward different regression functions?
6.9.20* Propose E-estimator for regression of Y on C. Explain its motivation, used probability formulas and assumptions.

6.11 Notes

Survival analysis is concerned with the inference about lifetimes, that is times to an event. The corresponding problems occur in practically all applied fields ranging from medicine, biology and public health to actuarial science, engineering and economics. A common feature of available data is that observations are modified by either censoring, or truncation, or both.

There is a vast array of books devoted to this topic, ranging from those using a mathematically nonrigorous approach to mathematically rigorous books using a wide range of theories including empirical processes, martingales in continuous time and stochastic integration among others. The literature is primarily devoted to parametric and semiparametric inference as well as nonparametric estimation of the survival function, and the interested reader can find many interesting examples, ad hoc procedures, advanced theoretical results and a discussion of using different software packages in the following books: Kalbfleisch and Prentice (2002), Klein and Moeschberger (2003), Martinussen and Scheike (2006), Aalen, Borgan and Gjessing (2008), Hosmer et al. (2008), Kosorok (2008), Allison (2010, 2014), Guo (2010), Fleming and Harrington (2011), Mills (2011), Royston and Lambert (2011), van Houwelingen and Putter (2011), Wienke (2011), Chen, Sun and Peace (2012), Crowder (2012), Kleinbaum and Klein (2012), Klugman, Panjer and Willmot (2012), Liu (2012), Lee and Wang (2013), Li and Ma (2013), Allison (2014), Collett (2014), Klein et al. (2014), Harrell (2015), Zhou (2015), Moore (2016), Tutz and Schmid (2016), and Ghosal and van der Vaart (2017).

6.1 Nonparametric estimation of the hazard rate is a familiar topic in the literature. Different type of estimators, including kernel, spline, classical orthogonal series and modern wavelet methods, have been proposed. A number of adaptive, to smoothness of an underlying hazard rate function, procedures motivated by known ones for the probability density, have been developed. A relevant discussion and thorough reviews may be found in a number of classical and more recent publications including Prakasa Rao (1983), Cox and Oakes (1984), Silverman (1986), Patil (1997), Wu and Wells (2003), Wang (2005), Gill (2006), Müller and Wang (2007), Fleming and Harrington (2011), Patil and Bagkavos (2012), Lu and Min (2014) and Daepp et al. (2015) where further references may be found. Interesting results, including both estimation and testing, have been obtained for the case of known restrictions on the shape of hazard rate, see a discussion in Jankowski and Wellner (2009). The plug-in estimation approach goes back to Watson and Leadbetter (1964), and see a discussion in Bickel and Doksum (2007). Boundary effect is a serious problem in nonparametric curve estimation. Complementing a trigonometric basis by polynomial functions is a standard method of dealing with boundary effects, and it is discussed in Efromovich (1999a, 2001a, 2018a,b).

The estimator $\hat{G}^X(x)$, defined in (6.1.14), is not equal to $1 - \hat{F}^X(x)$ where $\hat{F}^X(x) := n^{-1}\sum_{l=1}^{n} I(X_l \leq x)$ is the classical empirical cumulative distribution function. The reason for this is that we use reciprocal of $\hat{G}^X(X_l)$.

Efficient estimation of the hazard rate and the effect of the interval of estimation on the MISE is discussed in Efromovich (2016a, 2017). It is proved that the E-estimation methodology yields asymptotically sharp minimax estimation.

6.2–6.4 The topic of estimation of the hazard rate from indirect observations, created by left truncation and right censoring (LTRC), as well as estimation of the distribution, have received a great deal of attention in the statistical literature with the main emphasis on parametric models and the case of RC. See, for example, books by Cox and Oakes (1984), Cohen (1991), Anderson et al. (1993), Efromovich (1999a), Klein and Moeschberger (2003), Fleming and Harrington (2011), Lee and Wang (2013), Collet (2014), Harrell (2015), as well as papers by Uzunogullari and Wang (1992), Cao, Janssen and Veraverbeke (2005), Brunel and Comte (2008), Qian and Betensky (2014), Hagar and Dukic (2015), Shi, Chen and Zhou (2015), Bremhorsta and Lamberta (2016), Dai, Restaino and Wang (2016), Talamakrouni, Van Keilegom and El Ghouch (2016), and Wang et al. (2017), where further references may be found. Estimation of the change point is an interesting and related problem in survival analysis, see a review and discussion in Rabhi and Asgharian (2017). Bayesian approach is discussed in Ghosal and van der Vaart (2017) where further references may be found.

Efficiency of the proposed E-estimation methodology is established in Efromovich and Chu (2018a,b) where a numerical study and practical examples of the analysis of cancer data and the longevity in a retirement community may be found.

6.5 Nelson–Aalen estimator of the cumulative hazard is a popular choice in survival analysis. The estimator is defined as follows. Suppose that we observe right censored survival times of n patients meaning that for some patients we only know that their true survival times exceed certain censoring times. Let us denote by $X_1 < X_2 < \dots$ the times when deaths are observed. Then the Nelson–Aalen estimator for the cumulative hazard of X is

$$\check{H}^X(x) := \sum_{l:\, X_l \le x} (1/R_l), \tag{6.11.1}$$

where R_l is the number of patients at risk of death (that is alive and not censored) just prior to time X_l. Note that the estimator is nonparametric.

If we plug the Nelson-Aalen estimator in the formula $G^X(x) = \exp(-H^X(x))$ for the survival function, then the obtained estimator is referred to as Nelson-Aalen-Breslow estimator. The interested reader may compare this estimator with Kaplan-Meier estimator (6.5.3) and realize their similarity, the latter is also supported by asymptotic results. Further, estimator (6.5.10) is formally identical to the Nelson–Aalen–Breslow estimator while the underlying idea of its construction is based on the sample mean methodology. This remark sheds additional light on similar performance in the simulations of the sample mean and Kaplan–Meier estimators of survival function. Of course, there is a vast variety of different ideas and methods proposed in the literature, see the above-cited books as well as Woodroofe (1985), Dabrowska (1989), Antoniadis, Gregoire and Nason (1999), Efromovich (1999a, 2001a), De Una-Álvarez (2004), Wang (2005), Brunel and Comte (2008) and Wang et al. (2017).

Estimation under shape restrictions is an important part of survival analysis and special procedures are suggested for taking the restrictions into consideration, see a discussion in Groeneboom and Jongbloed (2014). See also Srivastava and Klassen (2016). For the E-estimation, there is no need to make any adjustment, instead, after calculating an E-estimate, it is sufficient to take a projection on the class of assumed functions and make the estimate bona fide. A theoretical justification of this approach can be found in Efromovich (2001a).

6.6 Numerical simulation is a popular statistical tool for a simultaneous analysis of several estimators, see a discussion in Efromovich (2001a) and Efromovich and Chu (2018a,b). A discussion of dependent data and further references may be found in El Ghouch and Van Keilegom (2008, 2009), Liang and de Una-Álvarez (2011) and De Una-Álvarez and Veraverbeke (2017).

6.7 An interesting extension of the discussed topic is to develop oracle inequalities for estimators under different loss functions. Here the approaches of Efromovich (2004e,f; 2007a,b) may be instrumental. Sequential estimation is another interesting topic, see Efromovich (2004d) where the case of direct observations is considered. See also Su and Wang (2012).

It is an interesting and open problem to develop a second-order efficient estimator of the survival function. Here the approach of Efromovich (2001b, 2004c) may be instrumental. Another area of developing is the efficient multivariate estimation, see a discussion in Harrell (2015) as well as corresponding results for direct observations in Efromovich (1999a, 2000b, 2010c). Specifically, an interesting approach would be to exploit the possibility of different smoothness of the density in variables. The latter is referred to as the case of an anisotropic distribution. Further, some of the variables may be discrete (the case of a mixed distribution), and this also may attenuate the curse of multidimensionality, see a discussion of the corresponding E-estmation methodology in Efromovich (2011c).

Bayesian approach may be also useful in the analysis of multivariate distributions, see Ghosal and van der Vaart (2017).

6.8–6.9 There is a number of interesting and practically important extensions of the considered problem. The first and natural one is to consider LTRC modifications. Sequential estimation with assigned risk is another natural setting, where similarly to Efromovich (2007d,e; 2008a,c) it is possible to consider estimation with assigned risk. Specific of the problem is that now censoring affects the risk, and the latter should be taken into account. Estimation of the conditional density is another important topic, here results of Efromovich (2007g, 2010b) can be helpful. See also Wang (1996), Delecroix, Lopez and Patilea (2008), Zhang and Zhou (2013), and Wang and Chan (2017).

Wavelet estimation is a popular choice for the nonparametric estimation, see a discussion in Wang (1995), Härdle et al. (1998), Mallat (1998), Efromovich (1999a), Vidakovic (1999), Nason (2008), Addison (2017), as well as an introductory text on wavelets by Nickolas (2017). The E-estimation methodology for wavelets and multiwavelets is justified in Efromovich (1999b, 2000a, 2001c, 2004e, 2007b), Tymes, Pereyra and Efromovich (2000), and Efromovich and Smirnova (2014a). Practical applications of the wavelet E-estimation are discussed in Efromovich et al. (2004), Efromovich et al. (2008), Efromovich (2009b), Efromovich and Smirnova (2014b).

Quantile regression is another traditionally studied statistical problem, see a discussion in Efromovich (1999a). Frumento and Bottai (2017) consider the problem of quantile regression under the LTRC where further references may be found. A tutorial on regression models for the analysis of multilevel survival data can be found in Austin (2017). Another related topic is the regression models for the restricted residual mean life, see a discussion in Cortese, Holmboe and Scheike (2017).