Deconvolving Kernel Density Estimation of Right Censored Duration Data with Recall Errors

T. K. Chakrabarty[*]

Department of Statistics, North Eastern Hill University, Shillong, India
*Corresponding author: tapankumarchakrabarty@gmail.com

Abstract Demographic and Health Surveys (DHS) collect information on several landmark events retrospectively from the life or birth histories and recollections of past events of individuals. Retrospective information of the sort is known to be affected by recall errors which result in the misplacement of dates, and the distortion of reports of duration. For example, the retrospectively reported ages of weaning for all births that occurred during the three or five years preceding the survey are right censored and commonly display marked heaping at durations 12, 18 and 24 months. The present article at first tries to understand whether the heaping is a result of true behavior and societal norms or it is an age dependent outcome. Further, under an additive error model, a kernel-type deconvolving density estimator of weaning time is proposed by smoothing the increments of Kaplan-Meier (KM) cumulative distribution function. Using simulated data it has been shown that in small and moderately censored samples these estimators can reduce the bias substantially. Finally, an empirical illustration is provided using National Family Health Survey (NFHS-3, 2005-06) data from India.

Keywords: weaning duration, survival distribution, kernel density, smoothing, deconvolution

1. Introduction

Duration data or time to event data occur in various fields of biological, engineering and social sciences. In demographic event history analysis, for instance, a primary focus has always been the timing and occurrence of events and consequences of these events for population stock. Data on the timing and date of events are often collected retrospectively [11] by interview, usually by conducting a survey or census for a wide variety of topics. Among others, these topics include marital events collected through marital histories in US Panel Study of Income Dynamics (PSID) [15], births from fertility histories in World Fertility Survey (WFS) [16], post-partum amenorrhea (the interval after a pregnancy before menstruation returns) in the Malaysian Family Life Surveys (MFLS) [4], maternal and child health and nutrition in Demographic and Health Surveys (DHS) [1,3].

The probability distribution of the time till occurrence of an event is useful for several important reasons. For instance, comparisons of decline in average durations of breastfeeding in developing countries and thereby recommending breastfeeding promotion high in their health policy agenda [18]; comparing the duration and stability of marital and non-marital unions in developed countries [6]; understanding of the factors which affect the timing and patterns of differential progression to the third birth [5,12] etc. Estimation of this distribution is an important inferential issue [10].

Distributions of event durations or times can be estimated from either current-status or retrospective history data [8]. Current-status data comprise information on whether the milestone has or has not been reached at the time of a survey. If the milestone has been reached, we have incomplete information on when this occurred. On the other hand, we do not know when it will be reached (if ever) for those respondents who have not achieved the milestone at the time of survey. Current –status data thus correspond to the extreme situation where all the survival time data are either right-censored or left-censored. While these data constraints are restrictive, it is still possible to estimate the distribution using parametric models [13,14], and to estimate some non-parametric models [10].

Retrospective data on the other hand, comprise information from the respondents that the milestone (1) occurred at a certain age or (2) has not yet occurred, i.e. right-censored data (age at milestone > age at survey). The major attraction of retrospective data is the obvious ease of collection and researchers often opt for cross-sectional studies in order to save time and cost. Despite their advantages, retrospectively reported data also have the drawback that people may not accurately report when events occur. It is a stylized fact that, when people report the timing of events that happened in the distant past, they tend to round up or down the year or time since the event occurred. Consequently, events tend to be "heaped" on multiples of chronological or calendar units. In case of age in months at weaning for instance, as we shall see in section 2 later, retrospectively reported ages of weaning

commonly display marked heaping at durations 6, 12, 18 and 24 months.

Various studies concerning the distribution of weaning times [8,18] have observed that the current-status measures lead to unbiased estimates of the survival function for a sample of births that occur during a fixed period. The estimates of the survival function so obtained, however, often fail to decline monotonically and are subject to larger variations. Retrospective history data are to be preferred because the sample size used to estimate the hazard rate at each duration is larger than is the case for current- status information, so that sampling variation is lower. If, however, the accuracy of the retrospective information is suspect, then the current-status data may be preferable despite the higher sampling variability. Thus, the method of analysis must be chosen by taking into account knowledge of the accuracy of a specific data set and an assessment of the extent of misreporting of duration information.

While dealing with the retrospectively reported right censored duration data, the present article proposes to examine whether recall errors and heaping of ages is universal. Whether such heaping is genuine for some obvious reasons such as societal norms or it is dependent on time since occurrence of the event. The main purpose of this article is to obtain probability density function of the right censored duration variable from the data contaminated with recall errors. The paper is organized as follows. In section 2, I discuss the sources of data used in the present work and some of the methodological issues in analyzing the information on breastfeeding from large scale health surveys. Kaplan-Meier and Nelson-Aalen estimates of survival functions are obtained in section 3 using data from the last two National Family Health Surveys for six north-eastern states in India. Following this, I propose a kernel-type deconvolving density estimator [2,19] of durations of breastfeeding under non-negligible additive type recall error and has been applied to the data mentioned above.

2. Data on Weaning Time

Demographic and Health Surveys (DHS) collect information on timing and age of several landmark events by a representative cross-sectional sample of ever married women and their children retrospectively from the life or birth histories and recollections of past events. For instance, respondent mothers are asked the following questions for each of her children born during the last three/five years before the survey:

a. Did you ever breastfed the child?
b. Are you still breastfeeding the child?
c. For how many months did you breastfeed the child?

The dichotomous answers ("yes" or "no") from the first two questions are known as current-status data. In addition, if respondents answer the age at which the child is weaned when the answer for (b) is a "no", it is called retrospective data. When the child is still breastfeeding at the time of survey, the number of months the child already breastfed is taken as a right censored observation for the weaning time.

For the present analysis, the information on breastfeeding in eight north-eastern states of India:

Arunachal Pradesh, Assam, Manipur, Meghalaya, Mizoram, Nagaland, Sikkim and Tripura have been pooled based on National Family Health Survey (NFHS-3) [7]. The NFHS-3, conducted during December 2005 to August 2006, gathered information on 21,843 ever-married women aged 15 – 49. Information on breastfeeding was collected for the children of interviewed women born in the five years preceding NFHS-3. For any given woman, a maximum of three births were included in the analysis of NFHS-3. For a total of 7600 children information on breastfeeding duration was available, of whom 3759 (49.5%) children were still breast feeding at the date of interview or were breastfed until died. The rest of the children (50.5%) completed their breast feeding. The duration of breastfeeding for the births who were still breastfeeding was calculated as the difference between their birth dates and the date of the survey. Table 1 below, reports the state wise distribution of number of children by the status of completeness of information.

Table 1. Statewise Distribution of Sample Data

State	No. of House holds	No. of women aged 15-49 interviewed	No. of Children		
			still breast feeding	weaned	Total
Arunachal Pradesh	1526	1647	365	321	686
Assam	3437	3840	728	487	1215
Manipur	3498	4512	983	646	1629
Meghalaya	1900	2124	261	390	651
Mizoram	1513	1791	272	459	731
Nagaland	3866	3896	586	1070	1656
Sikkim	1902	2127	235	255	490
Tripura	1574	1906	329	213	542
Total	19216	21843	3759	3841	7600

2.1. Recall Error and Its Characteristics

As mentioned earlier, retrospectively reported data have the drawback that people may not accurately report when events occur [15,18]. It is a stylized fact that, when people report the timing of events that happened in the distant past, they tend to round up or down the year or time since the event occurred. Consequently, events tend to be "heaped" on multiples of chronological or calendar units (e.g. on units of five or ten for data that naturally occur over years). In case of retrospectively reported ages of weaning, the data commonly display marked heaping at durations 6, 12, 18 and 24 months. We illustrate this by plotting the proportion of births for both the status category: children who are weaned and children who are still breastfed against their age in months in Figure 1.

Figure 1 demonstrates that heaping of ages at weaning is pronounced only for those children who are already weaned at the time of interview. At the same time, we also have the following hypothesis to be validated from the sample at hand:

H_{01}: Heaping of weaning time at multiple of 6 months is genuine and it is due to the existing social norms.

Secondly, people are notoriously poor at recalling events and the timing of events [11]. In general, more errors occur the greater is the time-lag between an event and its recall. Thus, the second hypothesis to be validated is as following:

H$_{02}$: For the weaned children, the greater is the time lag between weaning and its recall, the higher is the heaping.
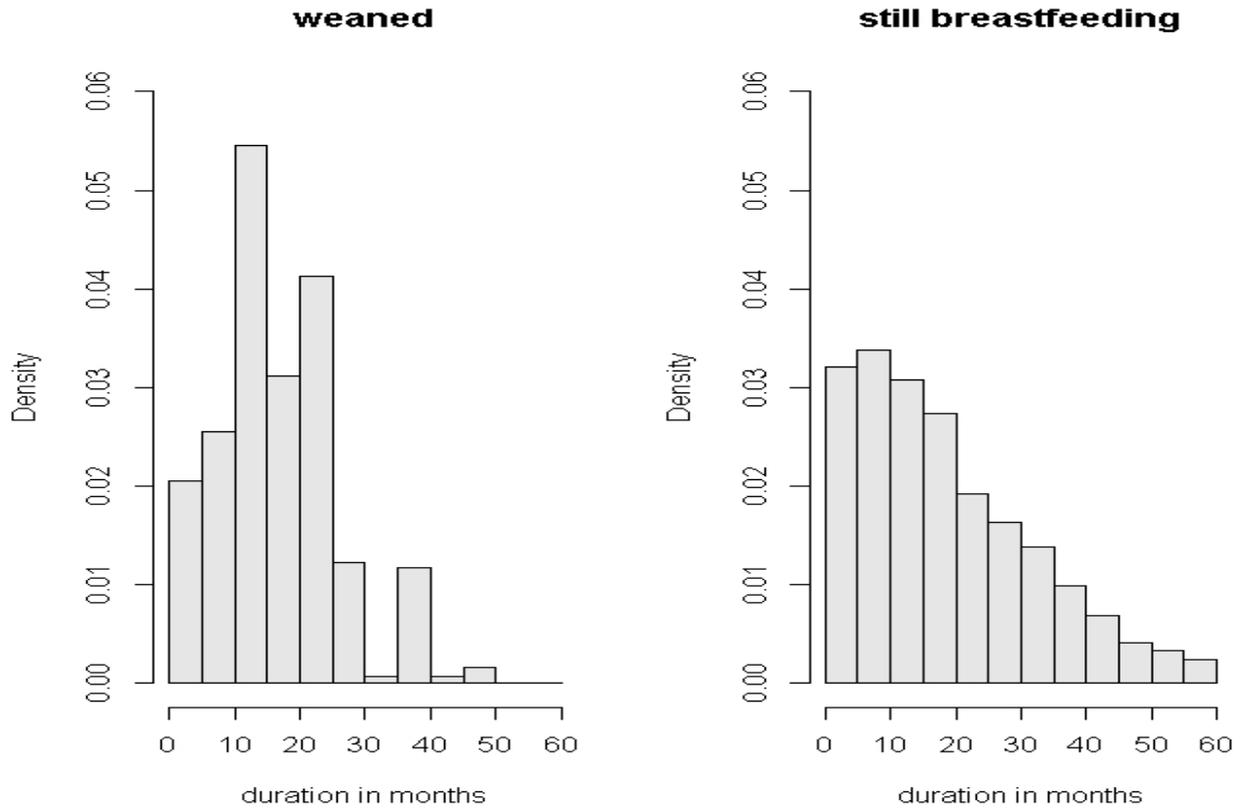
Figure 1. Histograms of Weaning durations for weaned and currently breastfeeding children

2.1.1. Is Heaping a Result of True Behaviour?

If heaping is resulted from a true behavior, such as, social norms of weaning at ages multiple of 6 months, denoted as T6, we would then find (i) substantially more masses at T6 than at T6-1 or T6 + 1, and (ii) significant differences between the masses at T6-1 and T6+1 months. Tab**l**e **2** below provides the probabilities of weaning at ages multiple of 6 months along with its preceding and succeeding values. Also, the p-values of testing the hypothesis:

$$\mathbf{H_0}: p_{I[T6-1]} = p_{I[T6+1]} \text{vs } H_1: p_{I[T6-1]} \neq p_{I[T6+1]}.$$

Table 2. Probability of weaning at ages multiple of 6 months

T6	P$_{I[T6]}$	P$_{I[T6-1]}$	P$_{I[T6+1]}$	p-value
6	p$_6$ =0.0058	p$_5$ =0.0034	p$_7$ =0.0069	0.1
12	p$_{12}$ =0.0855	p$_{11}$ =0.0081	p$_{13}$ =0.0184	0.01
18	p$_{18}$ =0.0808	p$_{17}$ =0.0113	p$_{19}$ =0.0103	0.84
24	p$_{24}$ =0.0955	p$_{23}$ =0.0036	p$_{25}$ =0.0076	0.42

Results are indicative of the fact that there is no strong evidence against the equality of proportions of weaning at ages T6-1 and T6+1. Moreover, if most of the children are weaned at ages T6 i.e. multiple of 6 months, we should have observed p$_{I[T6-1]}$ > p$_{I[T6+1]}$; which has not happened for the present sample.

2.1.2. Age Dependent Heaping of Weaning Times

As earlier, if T6 denote the age that is a multiple of 6 months, let I[T6] be an indicator of the reported weaning age is a multiple of 6 months; so that, I[T6] = 1, if the reported weaning time is a multiple of 6 months and 0 otherwise. Our objective here is to test the hypothesis H$_{02}$

as stated earlier. To formally carry this out, we fit the logistic regression model

$$\log\left(\frac{P_{I[T6]}}{1 - P_{I[T6]}}\right) = \beta_0 + \beta_1 \times currentAge$$

where 'currentAge' is the proxy covariate for the time lag between the timing of weaning and the date of interview at recall of the event. The results of the test

$$H_{02}: \beta_1 = 0 \quad H_{12}: \beta_1 \neq 0$$

is provided in Table 3. We find that the effect of this proxy covariate is highly significant.

Table 3. Estimate of the effect recall age on heaping

Coefficient	Estimate	Std Error	p-value
Intercept	-1.897	0.098	0.000
Age	0.047	0.004	0.000

3. Density Function Estimators

We consider non-parametric procedure for estimating the probability S(x) of surviving to time x, using a random sample X$_1$, X$_2$, …, X$_n$ of death times from a distribution F(x). The X$_i$ are censored on the right by random variables C$_i$, so that one observes only min(X$_i$, C$_i$) = Y$_i$, i=1,…..,n. The C$_i$'s constitute a random sample, drawn independently of the X$_i$, from a distribution G(c). We let Y$_{(1)} \leq$ ….$\leq$ Y$_{(n)}$ denote the ordered observations, and let $\delta_i = I\left(X_{(i)} \leq C_{(i)}\right)$ be an indicator for the event that Y$_{(i)}$ is uncensored. Kaplan and Meier [9] developed the nonparametric estimator of S(x) as

$$\hat{S}_{KM}(x) = \prod_{Y_{(i)} \leq x} \left(\frac{r_i - d_i}{r_i} \right)^{\delta_i} \tag{3.1}$$

where r_i = # alive at time $Y_{(i)}$-, d_i = # died at time $Y_{(i)}$.

Greenwood's formula for the variance of the survival function

$$\hat{V}(\hat{S}(x)) = (\hat{S}(x))^2 \sum_{Y_{(i)} \leq x} \frac{d_i}{r_i(r_i - d_i)}$$

$$or \quad \hat{V}(\ln \hat{S}(x)) = \sum_{Y_{(i)} \leq t} \frac{d_i}{r_i(r_i - d_i)} \tag{3.2}$$

The endpoints of a 100(1-α)% confidence interval for S(x) on the cumulative hazard or log-survival scale is given by

$$\exp(\ln \hat{S}(x) \pm z_{1-\alpha/2} s\hat{e}(\ln \hat{S}(x))) \tag{3.3}$$

Figure 2 shows the plots of Kaplan-Meier and Nelson-Aalen estimates of survival function and the jumps in these estimates may be observed at times multiple of 6 months as discussed earlier.
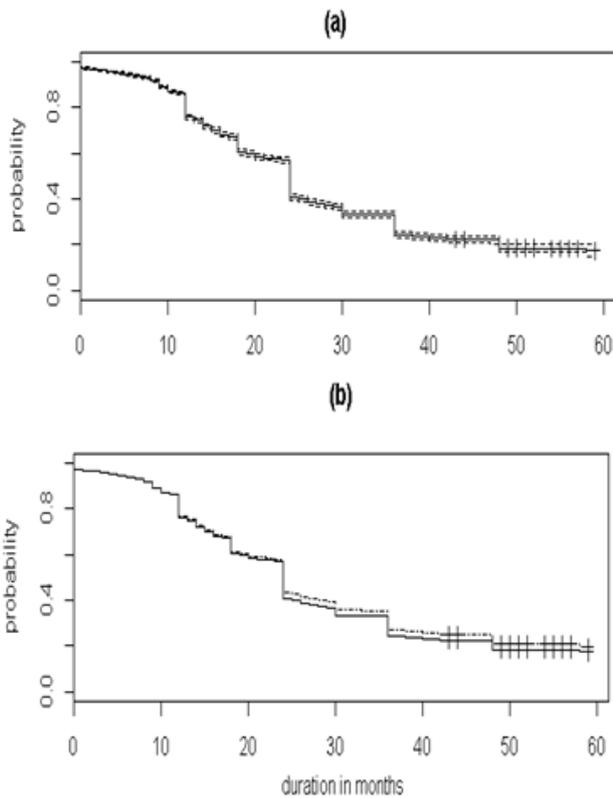
Figure 2. (a) Kaplan-Meier Survival Curve with 95% confidence band; (b) Comparison of K-M and Nelson –Aalen Estimates of Survival

A kernel density estimator of f_X can be motivated through the Kaplan-Meier estimator of the distribution function F_X, which is given by

$$\hat{F}_{KM}(x) = 1 - \hat{S}_{KM}(x)$$

$$= \begin{cases} 0 & , \quad 0 \leq x \leq Y_{(1)} \\ 1 - \prod_{Y_{(i)} \leq x} \left(\frac{r_i - d_i}{r_i} \right) & , \quad Y_{(i)} \leq x \leq Y_{(i+1)} \\ 1 & , \quad x > Y_{(n)} \end{cases}$$

The kernel estimator of $f_Y(x)$ induced by $\hat{F}_{KM}$ is then

$$\hat{f}_Y(x,h) = h^{-1} \int K\left(\frac{x-y}{h} \right) d\hat{F}_{KM}(y)$$

$$= h^{-1} \sum_j K\left(\frac{x-Y_{(j)}}{h} \right) s_j \tag{3.4}$$

where s_j is the size of the jump of $\hat{F}_{KM}$ at $Y_{(j)}$.

We now assume that the observable times Y_1, Y_2, ……,Y_n are contaminated with non-negligible recall error such that

$$W_i = Y_i + Z_i , \quad i = 1,......n$$

and, for each i, Z_i is a random variable that is independent of Y_i and has known density f_Z, which we call the error density. If we apply the ordinary kernel estimate to the W_1, …, W_n then we will obtain a consistent estimate of convolution

$$f_W = f_Y * f_Z$$

rather than f_Y which we aim to estimate. Estimation of f_Y requires that we take into account the fact that it is convolved with f_Z to give the density of the error contaminated data. Thus the estimation of f_Y is a problem of deconvolution type. A kernel type solution is obtained by using Fourier transform (or characteristic function) properties and noting that

$$\varphi_{f_W}(t) = \varphi_{f_Y}(t)\varphi_{f_Z}(t)$$

where ϕ_g is used to denote the c.f. of a density g. According to the Fourier inversion theorem, the target density can be written as

$$f_Y(y) = (2\pi)^{-1} \int e^{-ity} \varphi_{f_Y}(t) dt$$

$$= (2\pi)^{-1} \int e^{-ity} \left\{ \varphi_{f_W}(t)/\varphi_{f_Z}(t) \right\} dt$$

provided $\varphi_{f_Z}(t) \neq 0 \quad \forall t$. An estimate of $f_Y(y)$ is obtained by replacing f_W by its kernel estimator

$$\hat{f}_W(y,h) = (nh)^{-1} \sum_{i=1}^{n} K\left(\frac{y-W_i}{h} \right)$$

to obtain

$$\hat{f}_Y(y,h) = (2\pi)^{-1} \int e^{-ity} \left\{ \varphi_{\hat{f}_W(y,h)}(t)/\varphi_{f_Z}(t) \right\} dt$$

which is the deconvolving kernel density estimator [17]. It can be shown [19] that the deconvolving kernel density estimator of the target density is

$$\hat{f}_Y(y,h) = (nh)^{-1} \sum_{i=1}^{n} K^Z\left(\frac{y-W_i}{h}, h \right) \tag{3.5}$$

where

$$K^Z(u,h) = (2\pi)^{-1} \int e^{-ity} \left\{ \varphi_K(t)/\varphi_{f_Z}(t/h) \right\} dt \tag{3.6}$$

φ_K being the characteristic function of the kernel K used in estimating $\hat{f}_W$. Thus, the kernel $K^Z(\cdot, h)$ is to be used for estimating f_Y instead of K. This effective kernel differs

from K in that its shape depends on the bandwidth. We now use $K^Z(\cdot,h)$ of (3.6) to rewrite (3.4) as

$$\hat{f}_Y(x,h) = h^{-1}\sum_j K^Z\left(\frac{x-Y_{(j)}}{h},h\right)s_j. \qquad (3.7)$$

3.1. Simulation Study of Small sample Bias

We generate data through simulation to examine the small sample bias of the estimator $\hat{f}_Y(x,h)$ in (3.7). The effective kernel is obtained for two different error density functions. When the error density is Laplacian

$$f_Z(x) = (2\sigma)^{-1}\exp(-|x|/\sigma), \quad -\infty < x < \infty; \quad \sigma > 0$$

and that $K(x) = \Phi(x)$, the standard normal kernel, the effective kernel for deconvolution of Laplacian error is

$$K^Z(x,h) = \Phi(x)\left\{1+(\sigma/x)^2(x^2-1)\right\} \qquad (3.1.1)$$

Table 4. Small sample Bias B_0, B_L and B_N for three different sample sizes and three different censoring percentages at selected time points

n	λ_1	λ_2	π	Time	Bias		
					B_0	B_L	B_N
30	8	-	0		h = 1.8	h = 1.6, σ = 0.8	h = 2.2, σ = 0.9
				6	-0.009	-0.003	-0.006
				12	-0.002	0.000	-0.001
				18	-0.002	-0.001	-0.002
				24	0.000	0.000	-0.001
30	12	36	27		h = 2.0	h = 2.0, σ = 0.7	h = 2.5, σ = 1.0
				6	0.000	0.001	0.001
				12	-0.016	-0.009	-0.010
				18	0.009	0.013	0.012
				24	-0.013	-0.010	-0.010
				30	0.000	0.000	0.000
				36	0.000	0.000	0.000
30	14	12	53		h = 2.5	h = 2.5, σ = 1.0	h = 3.0, σ = 1.0
				6	0.015	0.014	0.014
				12	-0.025	-0.017	-0.020
				18	-0.003	-0.003	-0.003
				24	-0.001	-0.002	-0.002
60	18	-	0		h = 3.0	h = 3.0, σ = 1.0	h = 3.5, σ = 1.0
				6	0.006	0.005	0.005
				12	-0.009	-0.006	-0.006
				18	0.002	0.000	0.000
				24	-0.002	-0.001	-0.001
				30	0.002	0.001	0.001
				36	0.000	0.000	0.000
60	18	18	43		h = 2.5	h = 2.5, σ = 1.0	h = 3.5, σ = 1.0
				6	0.012	0.012	0.012
				12	-0.012	-0.006	-0.008
				18	0.001	-0.001	-0.001
				24	-0.005	-0.003	-0.004
				30	0.002	0.003	0.003
60	22	12	77		h = 2.0	h = 2.0, σ = 1.0	h = 2.5, σ = 1.0
				6	-0.002	0.005	0.001
				12	-0.005	0.001	-0.002
				18	0.001	-0.001	0.000
				24	0.002	-0.001	0.000
				30	0.000	0.000	0.000
100	8	-	0		h = 2.0	h = 2.0, σ = 0.8	h = 2.5, σ = 1.0
				6	-0.002	0.001	0.001
				12	0.000	0.000	0.000
				18	-0.004	-0.003	-0.003
				24	-0.001	-0.002	-0.002
				30	-0.003	-0.001	-0.001
				36	0.001	0.001	0.001
100	18	22	39		h = 2.0	h = 2.0, σ = 1.0	h = 2.5, σ = 1.0
				6	0.003	0.003	0.003
				12	-0.001	0.003	0.004
				18	0.001	0.001	0.000
				24	-0.005	-0.003	-0.004
				30	-0.001	0.000	0.000
100	22	18	60		h = 2.5	h = 2.5, σ = 1.0	h = 2.5, σ = 1.0
				6	-0.002	0.002	0.000
				12	0.005	0.004	0.004
				18	0.001	0.001	0.001
				24	0.003	0.004	0.003
				30	0.000	0.000	0.000

Supposing instead that the error variable has a N(0,σ²) distribution, the effective kernel would be

$$K^Z(x,h) = \frac{1}{\sqrt{2\pi\left(1-\left(\sigma/h\right)^2\right)}} \exp\left(-\frac{x^2}{2\left(1-\left(\sigma/h\right)^2\right)}\right). \quad (3.1.2)$$

Further, we use $F_X(x) = 1 - \exp(-x/\lambda_1)$ and $G_C(c) = 1 - \exp(-c/\lambda_2)$. The choices of pair of values for (λ_1, λ_2) give rise to desired censoring proportion to indicate no censoring, moderate censoring and heavy censoring. Observations have been simulated from F_X and G_C for three different sample sizes n=30, 60, 100. The pulling mechanism to contaminate the simulated data is defined through the following function

$$f(y) = \begin{cases} 6, & 4 \le y < 9 \\ 12, & 9 \le y < 15 \\ 18, & 15 \le y \le 20 \\ 24, & 21 \le y \le 27 \\ 30, & 28 \le y \le 32 \end{cases}$$

Table 4 presents the results showing bias at selected time points. We denote by B_0, B_L and B_N the bias due to the density estimate $\hat{f}_Y(x)$ in (3.4) without accounting for recall error, the bias due to the Laplacian error corrected density estimate and the bias due to the normal error corrected density estimate respectively. Table 4 illustrates the following general findings: (i) All estimators are fairly unbiased. (ii) The kernel density estimator $\hat{f}_Y(x)$ by smoothing the Kaplan-Meier distribution function.

3.2. Density Estimate from Duration of Breastfeeding Data

We now obtain smooth density estimates from duration of breastfeeding data of NFHS-3. Figure 3(a) provides smooth estimates of density of breastfeeding duration for two different choices of bandwidths (h = 2.5, 4.5). Figure 3(b) compares (i) smoothed estimate from Kaplan-Meier distribution function, (ii) estimate using Laplacian error correction, and (iii) estimate using normal error correction. It is well observed from Figure 3(a) that for a smaller choice of smoothing parameter (bandwidth = 2.5), the estimate of density is marked by the peaks at ages 12, 24, 36, and 48 months. A higher value (bandwidth = 4.5) however, is capable of smoothing the peaks which are thought to be not genuine. The right skewness of the distribution of weaning time is evident from the plots. This is quite reasonable as only few children are breastfed for longer durations. The median age at weaning is estimated to be 24 months with a 95% confidence interval (23.9, 24.1). Overall, an estimated proportion of about 60% children were continued their breastfeeding till the age of 24 months in the north eastern states of India.

The plots in Figure 3(b) show that there has been a locational shift in the density curve after the error correction. Both normal and laplacian error models try to shift the error corrected density towards the left. However, there is hardly any difference in the error corrected density due to the selection of the normal and laplacian error models.

The error corrected smooth density sharply rises reaching its maximum approximately at 12 months, followed by a plateau till 24 months and then trails off gradually. It describes well the situation in the whole population that more than three fourth of the children continue their breastfeeding till the age of 12 months. The long right tail is a result of the few subjects who had long breastfeeding experiences.
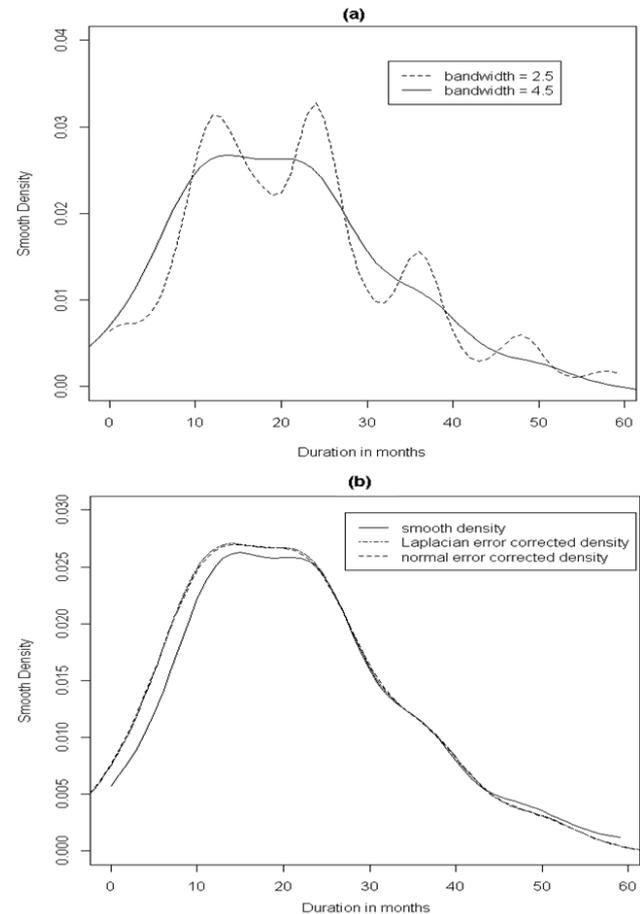
Figure 3. (a) Heaped and Smooth Density of Weaning Time; (b) Error Corrected Densit

4. Concluding Remarks

In this paper, a well-recognized feature of retrospectively collected right censored duration data on events that occur in the distant past - heaping on some natural time unit has been investigated. It is demonstrated that such heaping is pronounced for the subjects with completed outcome. Two important hypotheses characterizing this feature (1) genuineness and (2) time dependence has been validated. It has been observed that there is no evidence of heaping is genuine due to social norm and practice. Also, the pattern is age dependent i.e. the greater is the time lag of reporting, the more the heaping is.

Having verified the presence of non-negligible recall error in the data we have computed smooth error corrected density under an additive error model. The kernel estimator of density function for breastfeeding duration proposed in this article use the idea of convolving a kernel weight with the density estimates induced by the natural

estimate of the cumulative distribution function. The estimator has attractive mean squared error property [19] and is pointwise consistent [2].

For large samples, the effect of bias correction has been found to be little, subject to the appropriate selection of bandwidth parameter. One can make judicious use of any of the several hi-tech bandwidth selectors to implement the proposed estimator.

The methodology proposed in this article is of utmost practical use for the nations [18], where government intervention programmes are under process to increase the percentage of exclusive breastfeeders till six months and it is necessary to evaluate the efficacy of such programmes through bias free estimates of proportion of weaned cases.

References

[1]   Anandaiah, R. and Choe M. K. (2000). Are the WHO guidelines on Breastfeeding Appropriate for India? National Family Health Survey Subject Reports No. 16, International Institute for Population Sciences, Mumbai, India.

[2]   Anderson, P.K.,Borgan, Ø., Gill, R.D., and Keiding, N. (1993). Statistical Models Based on Counting Processes. Springer-Verlag New York, Inc.

[3]   Arnold, F., Parasuraman, S., Arokiasamy, P., and Kothari, M. (2009). *Nutrition in India*. National Family Health Survey (NFHS-3), India, 2005-06. Mumbai: International Institute for Population Sciences; Calverton, Maryland, USA: ICF Macro.

[4]   Beckett, M., Vanzo, J.D., Sastry, N., Panis, C.,and Pterson, C. (2001). The quality of retrospective data: an examination of long termrecall in a developing country. Journal of Human Resource, 36(3): 593-625.

[5]   Heckman, J.J., and Walker, J.R. (1992). Understanding third births in Sweden. In: Demographic applications of event history analysis, Trussell, J., Hankinson, R., and Tilton, J., (Eds.), Oxford University Press, Oxford, pp: 157-208.

[6]   Hoem, b. and Hoem, J. N. (1992). The disruption of marital and non-marital unions in contemporary Sweden. In: Demographic applications of event history analysis, Trussell, J., Hankinson, R., and Tilton, J.,(Eds.), Oxford University Press, Oxford, pp: 61-93.

[7]   International Institute for Population Sciences and ORC Macro. (2002). National Family Health Survey (NFHS-3), 2005-06: Northeastern States, IIPS, Bombay.

[8]   John, A.M., Menken, J,A., and Trussell, J. (1988). Estimating the distribution of interval lemgth: current status and retrospective history data. Population Studies 42: 115-127.

[9]   Kaplan, E.L. and Meier, P. (1958). Nonparametric estimation from incomplete observations, Journal of American Statistical Association, 53:457-481.

[10]   Mirzaei, S.S., and Sengupta, D. (2013). Parametric estimation of menarcheal age distribution based on recall data. Thecnical Report No. ASD/2013/3, Applied Statistical Unit, Indian Statistical Institute, 3. Available at www.isical.ac.in/asu/TR/TechRepASD201303.pdf.

[11]   Moss, L., and Goldstein, H. (1985). (eds.). The Recall Method in Social Surveys, Heinemann, Portsmouth.

[12]   Murphy, M. (1992). The progression to the third birth in Sweden. In: Demographic applications of event history analysis, Trussell, J., Hankinson, R., and Tilton, J.,(Eds.), Oxford University Press, Oxford, pp: 141-156.

[13]   Nelson, W. (1978). Life Data Analysis for Units Inspected for Failure (Quantal Response Data). IEEE Transactions on Reliability, R-27, 274-279.

[14]   Nelson, W. (1982). Applied Life data Analysis. John Wiley & Sons, New York.

[15]   Peters, H. E. (1988). "Retrospective Versus Panel Data in Analyzing Lifecycle Events." Journal of Human Resources 23(4): 488-513.

[16]   Smith, D.J. and Ferry, B. (1984). Correlates of Breastfeeding: World Fertility Survey, Comparative Studies, 41, Int. Stat. Inst., Voorburg.

[17]   Stefanski, L. and Carroll, R.J. (1990) Deconvolving kernel density estimators. Statistics 2, 169-184.

[18]   Trussell, J., Strawn, L.G., Rodriguez, G., and Vanlandingham, M. (1992). Trends and differentials in breastfeeding behavior: Evidence from WFS and DHS, Population Studies, 46:285-307.

[19]   Wand, M.P., and Jones, M.C. (1995). Kernel Smoothing. Chapmann & Hall/CRC, www.crcpress.com.