4.6 Estimators of the Survival Function for Left-Truncated and Right-Censored Data

The estimators and confidence intervals presented in sections 4.2–4.5 were based on right-censored samples. In this section, we shall show how these statistics can be modified to handle left-truncated and right-censored data. Here, we have associated, with the jth individual, a random age L_j at which he/she enters the study and a time T_j at which he/she either dies or is censored. As in the case of right-censored data, define $t_1 < t_2 < \cdots < t_D$ as the distinct death times and let d_i be the number of individuals who experience the event of interest at time t_i. The remaining quantity needed to compute the statistics in the previous sections is the number of individuals who are at risk of experiencing the event of interest at time t_i, namely Y_i. For right-censored data, this quantity was the number of individuals on study at time 0 with a study time of at least t_i. For left-truncated data, we redefine Y_i as the number of individuals who entered the study prior to time t_i and who have a study time of at least t_i, that is, Y_i is the number of individuals with $L_j < t_i \leq T_j$.

Using Y_i as redefined for left-truncated data, all of the estimation procedures defined in sections 4.2–4.4 are now applicable. However, one must take care in interpreting these statistics. For example, the Product-Limit estimator of the survival function at a time t is now an estimator of the probability of survival beyond t, conditional on survival to the smallest of the entry times $L, Pr[X > t \mid X \geq L] = S(t)/S(L)$. Similarly the Nelson–Aalen statistic estimates the integral of the hazard rate over the interval L to t. Note that the slope of the Nelson–Aalen estimator still provides an estimator of the unconditional hazard rate.

Some care in directly applying these estimators is needed. For left-truncated data, it is possible for the number at risk to be quite small for small values of t_i. If, for some t_i, Y_i and d_i are equal, then, the Product-Limit estimator will be zero for all t beyond this point, even though we are observing survivors and deaths beyond this point. In such cases, it is common to estimate the survival function conditional on survival to a time where this will not happen by considering only those death times beyond this point. This is illustrated in the following example.

EXAMPLE 4.3

To illustrate how the statistics developed in the previous sections can be applied to left-truncated data, consider the Channing House data described in section 1.16. The data is found in Table D.5 of Appendix D. Here the truncation times are the ages, in months, at which individuals

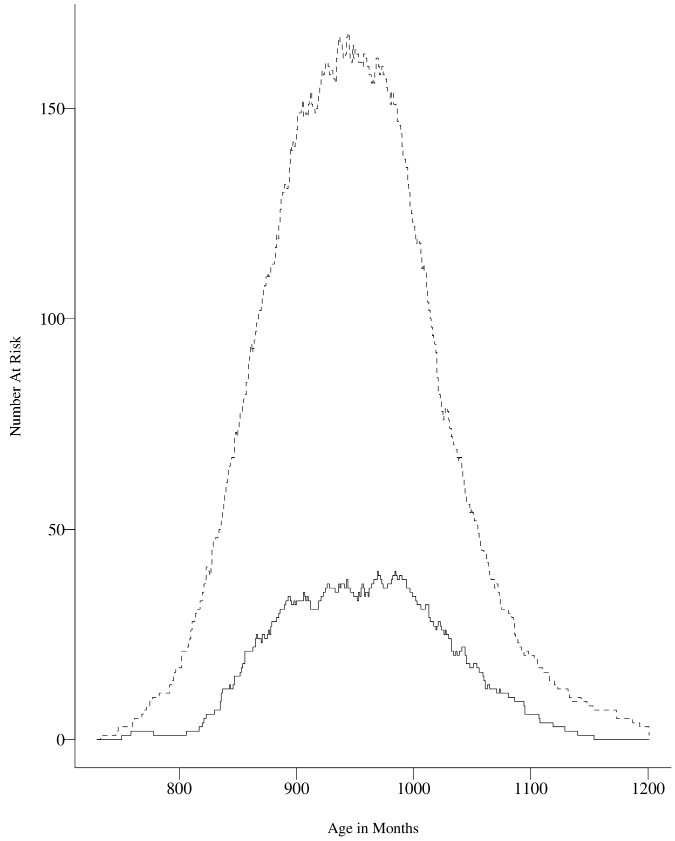

Figure 4.10 *Number at risk as a function of age for the 97 males (———) and the 365 females (-----) in the Channing house data set*

entered the community. We shall focus on estimating the conditional survival function.

Figure 4.10 shows the number of individuals at risk as a function of the age at which individuals die for both males and females. Note that the number at risk initially increases as more individuals enter into the study cohort and that this number decreases for later ages as individuals die or are censored.

Consider the data on males. Here the risk set is empty until 751 months when one individual enters the risk set. At 759 months, a second individual enters the risk set. These two individuals die at 777 and 781 months. A third individual enters the risk set at 782 months. Computing the Product-Limit estimator of $S(t)$ directly by (4.2.1) based on this data would yield an estimate of $\hat{S}(t) = 1$ for $t < 777$, $\hat{S}(t) = 1/2$ for $777 \leq t < 781$, and $\hat{S}(t) = 0$ for $t \geq 781$. This estimate has little meaning since the majority of the males in the study clearly survive beyond 781 months.

Rather than estimating the unconditional survival function, we estimate the conditional probability of surviving beyond age t, given survival to age a. We estimate $\boxed{S_a(t) = Pr[X > t \mid X \geq a]}$ by considering only those deaths that occur after age a, that is,

$$\hat{S}_a(t) = \prod_{a \leq t_i \leq t} \left[1 - \frac{d_i}{Y_i} \right], t \geq a. \tag{4.6.1}$$

Similarly for Greenwood's formula (4.2.2) or for the Nelson–Aalen estimator (4.2.3), only deaths beyond a are considered.

Figure 4.11 shows the estimated probability of surviving beyond age t, given survival to 68 or 80 years for both males and females.

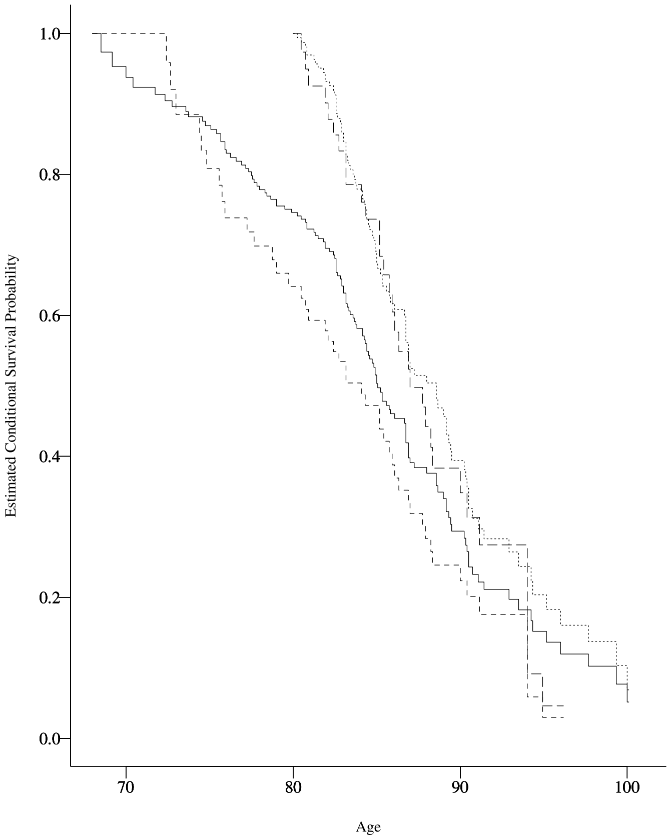

Figure 4.11 *Estimated conditional survival functions for Channing house residents. 68 year old females (————); 80 year old females (------); 68 year old males (— — —); 80 year old males (—— ——).*