# Nonlife Actuarial Models

## Chapter 10

## Model Estimation and Types of Data

# Learning Objectives

1. Parametric versus nonparametric estimation

2. Point estimate and interval estimate

3. Unbiasedness, consistency and efficiency

4. Failure-time data and loss data

5. Complete versus incomplete data, left truncation and right censoring

6. Individual versus grouped data

# 10.1 Estimation

### 10.1.1 Parametric and Nonparametric Estimation

- In the parametric approach, the distribution is determined by a finite number of parameters.

- Thus, the loss random variable $X$ has df $F(x;\theta)$ and pdf (pf) $f(x;\theta)$, where $\theta$ is the parameter of the df and pdf (pf).

- When $\theta$ is known, the distribution of $X$ is completely specified.

- In practical situations $\theta$ is unknown and has to be estimated using observed data. We denote $\hat{\theta}$ as an estimator of $\theta$ using the random sample.

- $F(x)$ and $f(x)$ may also be estimated directly for all values of $x$ without assuming specific parametric forms, resulting in nonparametric estimates of these functions.

### 10.1.2 Point and Interval Estimation

- As $\hat{\theta}$ assigns a specific value to $\theta$ based on the sample, it is called a **point estimator**.

- In contrast, an **interval estimator** of an unknown parameter is a random interval constructed from the sample data, which covers the true value of $\theta$ with a certain probability.

- Specifically, let $\hat{\theta}_L$ and $\hat{\theta}_U$ be functions of the sample data $\{X_1, \cdots, X_n\}$, with $\hat{\theta}_L < \hat{\theta}_U$ . The interval $(\hat{\theta}_L, \hat{\theta}_U)$ is said to be a $100(1-\alpha)\%$ **confidence interval** of $\theta$ if

$$\Pr(\hat{\theta}_L \leq \theta \leq \hat{\theta}_U) \geq 1 - \alpha. \tag{10.1}$$

### 10.1.3 Properties of Estimators

- As there are possibly many different estimators for the same parameter, an intelligent choice among them is important.

- We desire the estimate to be *close* to the true parameter value *on average*, leading to the unbiasedness criterion as follows.

**Definition 10.1 (Unbiasedness):** An estimator of $\theta$, $\hat{\theta}$, is said to be unbiased if and only if $\mathrm{E}(\hat{\theta}) = \theta$.

- In some applications, although $\mathrm{E}(\hat{\theta})$ may not be equal to $\theta$ in finite samples, it may approach to $\theta$ arbitrarily closely in large samples. We say $\hat{\theta}$ is **asymptotically unbiased** for $\theta$ if

$$\lim_{n\to\infty} \mathrm{E}(\hat{\theta}) = \theta. \tag{10.3}$$

- If we have two unbiased estimators, the closeness requirement suggests that the one with the smaller variance should be preferred. This leads us to the following definition.

**Definition 10.2 (Minimum Variance Unbiased Estimator):** Suppose $\hat{\theta}$ and $\tilde{\theta}$ are two unbiased estimators of $\theta$, $\hat{\theta}$ is more efficient than $\tilde{\theta}$ if $\mathrm{Var}(\hat{\theta}) < \mathrm{Var}(\tilde{\theta})$. In particular, if the variance of $\hat{\theta}$ is smaller than the variance of any other unbiased estimator of $\theta$, then $\hat{\theta}$ is the minimum variance unbiased estimator of $\theta$.

- While asymptotic unbiasedness requires the *mean* of $\hat{\theta}$ to approach $\theta$ arbitrarily closely in large samples, a stronger condition is to require $\hat{\theta}$ itself to approach $\theta$ arbitrarily closely in large samples. This leads us to the property of consistency.

**Definition 10.3 (Consistency):** $\hat{\theta}$ is a consistent estimator of $\theta$ if it **converges in probability** to $\theta$, which means that for any $\delta > 0$,

$$\lim_{n\to\infty} \Pr(|\hat{\theta} - \theta| < \delta) = 1. \tag{10.4}$$

- Note that unbiasedness is a property that refers to samples of all sizes, large or small. In contrast, consistency is a property that refers to large samples only.

**Theorem 10.1:** $\hat{\theta}$ is a consistent estimator of $\theta$ if it is asymptotically unbiased and $\text{Var}(\hat{\theta}) \to 0$ when $n \to \infty$.

- Biased estimators are not necessarily inferior if their average deviation from the true parameter value is small.

- We may use the **mean squared error** as a criterion for selecting estimators. The mean squared error of $\hat{\theta}$ as an estimator of $\theta$, denoted by $\mathrm{MSE}(\hat{\theta})$, is defined as

$$\mathrm{MSE}(\hat{\theta}) = \mathrm{E}[(\hat{\theta} - \theta)^2]. \tag{10.5}$$

We note that

$$\begin{aligned}
\mathrm{MSE}(\hat{\theta}) &= \mathrm{E}[(\hat{\theta} - \theta)^2] \\
&= \mathrm{E}[\{(\hat{\theta} - \mathrm{E}(\hat{\theta})) + (\mathrm{E}(\hat{\theta}) - \theta)\}^2] \\
&= \mathrm{E}[\{\hat{\theta} - \mathrm{E}(\hat{\theta})\}^2] + [\mathrm{E}(\hat{\theta}) - \theta]^2 + 2[\mathrm{E}(\hat{\theta}) - \theta]\mathrm{E}[\hat{\theta} - \mathrm{E}(\hat{\theta})] \\
&= \mathrm{Var}(\hat{\theta}) + [\mathrm{bias}(\hat{\theta})]^2. \qquad (10.6)
\end{aligned}$$

- $\mathrm{MSE}(\hat{\theta})$ is the sum of the variance of $\hat{\theta}$ and the squared bias. A small bias in $\hat{\theta}$ may be tolerated, if the variance of $\hat{\theta}$ is small so that the overall MSE is low.

**Example 10.2:** Let $\{X_1, \cdots, X_n\}$ be a random sample of $X$ which is distributed as $\mathcal{U}(0, \theta)$. Define $Y = \max\,\{X_1, \cdots, X_n\}$, which is used as an estimator of $\theta$. Calculate the mean, variance and mean squared error of $Y$. Is $Y$ a consistent estimator of $\theta$?

**Solution:** We first determine the distribution of $Y$. The df of $Y$ is

$$\begin{aligned}F_Y(y) &= \Pr(Y \leq y)\\ &= \Pr(X_1 \leq y, \cdots, X_n \leq y)\\ &= [\Pr(X \leq y)]^n\\ &= \left(\frac{y}{\theta}\right)^n.\end{aligned}$$

Thus, the pdf of $Y$ is

$$f_Y(y) = \frac{dF_Y(y)}{dy} = \frac{ny^{n-1}}{\theta^n}.$$

Hence, the first two raw moments of $Y$ are

$$\mathrm{E}(Y) = \frac{n}{\theta^n}\int_0^\theta y^n\,dy = \frac{n\theta}{n+1},$$

and

$$\mathrm{E}(Y^2) = \frac{n}{\theta^n} \int_0^\theta y^{n+1}\, dy = \frac{n\theta^2}{n+2}.$$

The bias of $Y$ is

$$\mathrm{bias}(Y) = \mathrm{E}(Y) - \theta = \frac{n\theta}{n+1} - \theta = -\frac{\theta}{n+1},$$

so that $Y$ is downward biased for $\theta$. However, as bias($Y$) tends to 0 when $n$ tends to $\infty$, $Y$ is asymptotically unbiased for $\theta$. The variance of $Y$ is

$$\begin{aligned} \mathrm{Var}(Y) &= \mathrm{E}(Y^2) - [\mathrm{E}(Y)]^2 \\ &= \frac{n\theta^2}{(n+2)(n+1)^2}, \end{aligned}$$

which tends to 0 when $n$ tends to $\infty$. Thus, by Theorem 10.1, $Y$ is a consistent estimator of $\theta$. Finally, the MSE of $Y$ is

$$\mathrm{MSE}(Y) = \mathrm{Var}(Y) + [\mathrm{bias}(Y)]^2$$

$$= \frac{2\theta^2}{(n+2)(n+1)},$$

which also tends to 0 when $n$ tends to $\infty$. □

# 10.2 Types of Data

## 10.2.1 Duration Data and Loss Data

- There are problems for which duration is the key variable of interest.

- Examples are: (a) the duration of unemployment of an individual in the labor force, (b) the duration of stay of a patient in a hospital, and (c) the survival time of a patient after a major operation.

- Depending on the specific problem of interest, the methodology may be applied to **failure-time data**, **age-at-death data, survival-time data** or any duration data in general.

- In nonlife actuarial risks a key variable of interest is the claim-severity or loss distribution.

- Examples of applications are: (a) the distribution of medical-cost claims in a health insurance policy, (b) the distribution of car insurance claims, and (c) the distribution of compensations of work accidents. These cases involve analysis of **loss data**.

### 10.2.2 Complete Individual Data

- Assume researcher has complete knowledge about the relevant duration or loss data of the individuals.

- Let $X$ denote the variable of interest (duration or loss), and $X_1, \cdots, X_n$ denote the values of $X$ for $n$ individuals.

- We denote the observed sample values by $x_1, \cdots, x_n$.

- There may be duplications of values in the sample, and we assume there are $m$ distinct values arranged in the order $0 < y_1 < \cdots < y_m$, with $m \leq n$.

- We assume $y_j$ occurs $w_j$ times in the sample, for $j = 1, \cdots, m$. Thus, $\sum_{j=1}^{m} w_j = n$.

- In the case of age-at-death data, $w_j$ individuals die at age $y_j$. If all individuals are observed from birth until they die, we have a **complete individual** data set.

- We define $r_j$ as the **risk set** at *time* $y_j$, which is the number of individuals in the sample exposed to the possibility of death at time $y_j$ (prior to observing the deaths at $y_j$).

- For example, $r_1 = n$, as all individuals in the sample are exposed to the risk of death just prior to time $y_1$.

- Similarly, we can see that $r_j = \sum_{i=j}^{m} w_i$, which is the number of individuals who are surviving just prior to time $y_j$.

**Example 10.3:** Let $x_1, \cdots, x_{16}$ be a sample of failure times of a machine part. The values of $x_i$, arranged in increasing order, are as follows

$$2,\ 3,\ 5,\ 5,\ 5,\ 6,\ 6,\ 8,\ 8,\ 8,\ 12,\ 14,\ 18,\ 18,\ 24,\ 24.$$

Summarize the data in terms of the set-up above.

**Solution:** There are 9 distinct values of failure time in this data set, so that $m = 9$. Table 10.1 summarizes the data in the notations described above.

**Table 10.1:** Failure-time data in Example 10.3

| $j$ | $y_j$ | $w_j$ | $r_j$ |
|---|---|---|---|
| 1 | 2 | 1 | 16 |
| 2 | 3 | 1 | 15 |
| 3 | 5 | 3 | 14 |
| 4 | 6 | 2 | 11 |
| 5 | 8 | 3 | 9 |
| 6 | 12 | 1 | 6 |
| 7 | 14 | 1 | 5 |
| 8 | 18 | 2 | 4 |
| 9 | 24 | 2 | 2 |

From the table it is obvious that $r_{j+1} = r_j - w_j$ for $j = 1, \cdots, m-1$. □

**Example 10.4:** Let $x_1, \cdots, x_{20}$ be a sample of claims of a group medical insurance policy. The values of $x_i$, arranged in increasing order, are as follows

15, 16, 16, 16, 20, 21, 24, 24, 24, 28, 28, 34, 35, 36, 36, 36, 40, 40, 48, 50.

There are no deductible and policy limit. Summarize the data in terms of the set-up above.

**Solution:** There are 12 distinct values of claim costs in this data set, so that $m = 12$. As there are no deductible and policy limit, the observations are ground-up losses with no censoring nor truncation. Thus, we have a complete individual data set. Table 10.2 summarizes the data in the notations described above.

**Table 10.2:** Medical claims data in Example 10.4

| $j$ | $y_j$ | $w_j$ | $r_j$ |
|---|---|---|---|
| 1 | 15 | 1 | 20 |
| 2 | 16 | 3 | 19 |
| 3 | 20 | 1 | 16 |
| 4 | 21 | 1 | 15 |
| 5 | 24 | 3 | 14 |
| 6 | 28 | 2 | 11 |
| 7 | 34 | 1 | 9 |
| 8 | 35 | 1 | 8 |
| 9 | 36 | 3 | 7 |
| 10 | 40 | 2 | 4 |
| 11 | 48 | 1 | 2 |
| 12 | 50 | 1 | 1 |

### 10.2.3 Incomplete Individual Data

- In certain studies the researcher may not have complete information about each individual observed in the sample.

- Consider a study on the survival time of patients after a surgical operation.

- When the study begins it includes data of patients who have recently received an operation. New patients who are operated during the study are included in the sample as well when they are operated. All patients are observed until the end of the study, and their survival times are recorded.

- If a patient received an operation some time before the study began, the researcher has the information about how long this patient has survived after the operation and the future survival time is conditional on this information.

- Other patients who received operations at the same time as this individual but did not live till the study began would not be in the sample.

- Thus, this individual is observed from a population which has been **left truncated,** i.e., information is not available for patients who do not survive till the beginning of the study.

- On the other hand, if an individual survives until the end of the study, the researcher knows the survival time of the patient up to that time, but has no information about when the patient dies.

- Thus, the observation pertaining to this individual is **right censored**, i.e., the researcher has the partial information that this individual's survival time goes beyond the study but does not know its exact value.

- We now define further notations for analyzing incomplete data. Using survival-time studies for exposition, we use $d_i$ to denote the left-truncation status of individual $i$ in the sample. Specifically, $d_i = 0$ if there is no left truncation (the operation was done during the study period), and $d_i > 0$ if there is left truncation (the operation was done $d_i$ periods before the study began).

- Let $x_i$ denote the survival time (time till death after operation) of the $i$th individual. If an individual $i$ survives at the end of the study, $x_i$ is not observed and we denote the survival time up to that time by $u_i$.

- Thus, for each individual $i$, there is a $x_i$ value or $u_i$ value (but not both) associated with it. The example below illustrates the construction of the variables introduced.

**Example 10.5:** A sample of 10 patients receiving a major operation is available. The data are collected over 12 weeks and are summarized in Table 10.3. Column 2 gives the time when the individual was first observed, with a value of zero indicating that the individual was first observed when the study began. A nonzero value gives the time when the operation was done, which is also the time when the individual was first observed. For cases in which the operation was done prior to the beginning of the study Column 3 gives the duration from the operation to the beginning of the study. Column 4 presents the time when the observation ceased, either due to death of patient (D in Column 5) or end of study (S in Column 5).

**Table 10.3:** Survival time after a surgical operation

| Ind $i$ | Time ind $i$ first obs | Time since operation when ind $i$ first obs | Time when ind $i$ ends | Status when ind $i$ ends |
|---|---|---|---|---|
| 1 | 0 | 2 | 7 | D |
| 2 | 0 | 4 | 4 | D |
| 3 | 2 | 0 | 9 | D |
| 4 | 4 | 0 | 10 | D |
| 5 | 5 | 0 | 12 | S |
| 6 | 7 | 0 | 12 | S |
| 7 | 0 | 2 | 12 | S |
| 8 | 0 | 6 | 12 | S |
| 9 | 8 | 0 | 12 | S |
| 10 | 9 | 0 | 11 | D |

Determine the $d_i$, $x_i$ and $u_i$ values of each individual.

**Solution:** The data are reconstructed in Table 10.4.

**Table 10.4:** Reconstruction of Table 10.3

| $i$ | $d_i$ | $x_i$ | $u_i$ |
|---|---|---|---|
| 1 | 2 | 9 | – |
| 2 | 4 | 8 | – |
| 3 | 0 | 7 | – |
| 4 | 0 | 6 | – |
| 5 | 0 | – | 7 |
| 6 | 0 | – | 5 |
| 7 | 2 | – | 14 |
| 8 | 6 | – | 18 |
| 9 | 0 | – | 4 |
| 10 | 0 | 2 | – |

- As in the case of a complete data set, we assume that there are $m$ distinct failure-time numbers $x_i$ in the sample, arranged in increasing order, as $0 < y_1 < \cdots < y_m$, with $m \leq n$.

- Assume $y_j$ occurs $w_j$ times in the sample, for $j = 1, \cdots, m$.

- Again, we denote $r_j$ as the **risk set** at $y_j$, which is the number of individuals in the sample exposed to the possibility of death at time $y_j$ (prior to observing the deaths at $y_j$).

- To update the risk set $r_j$ after knowing the number of deaths at time $y_{j-1}$, we use the following formula

$$\begin{aligned} r_j &= r_{j-1} - w_{j-1} + \text{number of observations with } y_{j-1} \leq d_i < y_j \\ &\quad - \text{number of observations with } y_{j-1} \leq u_i < y_j, \quad j = 2, \cdots, m. \end{aligned} \tag{10.8}$$

- Upon $w_{j-1}$ deaths at failure-time $y_{j-1}$, the risk set is reduced to $r_{j-1} - w_{j-1}$.

- This number is supplemented by the number of individuals with $y_{j-1} \leq d_i < y_j$, who are now exposed to risk at failure-time $y_j$, but are formerly not in the risk set due to left truncation.

- If an individual has a $d$ value that ties with $y_j$, this individual is *not included* in the risk set $r_j$.

- The risk set is reduced by the number of individuals with $y_{j-1} \leq u_i < y_j$, i.e., those whose failure times are not observed due to right censoring.

- If an individual has a $u$ value that ties with $y_j$, this individual is *not excluded* from the risk set $r_j$.

- Equation (10.8) can also be computed equivalently using the following formula

$$\begin{aligned} r_j &= \text{number of observations with } d_i < y_j - (\text{number of observations} \\ &\qquad \text{with } x_i < y_j + \text{number of observations with } u_i < y_j), \\ &\qquad\qquad \text{for } j = 1, \cdots, m. \end{aligned} \tag{10.9}$$

- Note that the number of observations with $d_i < y_j$ is the total number of individuals who are potentially facing the risk of death at failure-time $y_j$.

- Individuals with $x_i < y_j$ or $u_i < y_j$ are removed from this risk set as they have either died prior to time $y_j$ (when $x_i < y_j$) or have been censored from the study (when $u_i < y_j$).

- To compute $r_j$ using equation (10.8), we need to calculate $r_1$ using equation (10.9) to begin the recursion.

**Example 10.6:** Using the data in Example 10.5, determine the risk set at each failure time in the sample.

**Solution:** The results are summarized in Table 10.5. Columns 5 and 6 describe the computation of the risk sets using equations (10.8) and (10.9), respectively.

**Table 10.5:** Ordered death times and risk sets of Example 10.6

| $j$ | $y_j$ | $w_j$ | $r_j$ | Eq (10.8) | Eq (10.9) |
|---|---|---|---|---|---|
| 1 | 2 | 1 | 6 | — | $6-0-0$ |
| 2 | 6 | 1 | 6 | $6-1+3-2$ | $9-1-2$ |
| 3 | 7 | 1 | 6 | $6-1+1-0$ | $10-2-2$ |
| 4 | 8 | 1 | 4 | $6-1+0-1$ | $10-3-3$ |
| 5 | 9 | 1 | 3 | $4-1+0-0$ | $10-4-3$ |

It can be seen that equations (10.8) and (10.9) give the same answers. □

**Example 10.7:** Table 10.6 summarizes the loss claims of 20 insurance policies, numbered by $i$, with $d_i$ = deductible, $x_i$ = ground-up loss, and $u_i^*$ = maximum covered loss. For policies with losses larger than $u_i^*$, only the $u_i^*$ value is recorded. The right-censoring variable is denoted by $u_i$. Determine the risk set $r_j$ of each distinct loss value $y_j$.

**Table 10.6:** Insurance claims data of Example 10.7

| $i$ | $d_i$ | $x_i$ | $u_i^*$ | $u_i$ | $i$ | $d_i$ | $x_i$ | $u_i^*$ | $u_i$ |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 12 | 15 | – | 11 | 3 | 14 | 15 | – |
| 2 | 0 | 10 | 15 | – | 12 | 3 | – | 15 | 15 |
| 3 | 0 | 8 | 12 | – | 13 | 3 | 12 | 18 | – |
| 4 | 0 | – | 12 | 12 | 14 | 4 | 15 | 18 | – |
| 5 | 0 | – | 15 | 15 | 15 | 4 | – | 18 | 18 |
| 6 | 2 | 13 | 15 | – | 16 | 4 | 8 | 18 | – |
| 7 | 2 | 10 | 12 | – | 17 | 4 | – | 15 | 15 |
| 8 | 2 | 9 | 15 | – | 18 | 5 | – | 20 | 20 |
| 9 | 2 | – | 18 | 18 | 19 | 5 | 18 | 20 | – |
| 10 | 3 | 6 | 12 | – | 20 | 5 | 8 | 20 | – |

**Solution:** The distinct values of $x_i$, arranged in order, are

$$6, 8, 9, 10, 12, 13, 14, 15, 18,$$

so that $m = 9$. The results are summarized in Table 10.7. As in Table 10.5 of Example 10.6, Columns 5 and 6 describe the computation of the risk sets using equations (10.8) and (10.9), respectively.

**Table 10.7:** Ordered claim losses and risk sets of Example 10.7

| $j$ | $y_j$ | $w_j$ | $r_j$ | Eq (10.8) | Eq (10.9) |
|---|---|---|---|---|---|
| 1 | 6 | 1 | 20 | — | $20 - 0 - 0$ |
| 2 | 8 | 3 | 19 | $20 - 1 + 0 - 0$ | $20 - 1 - 0$ |
| 3 | 9 | 1 | 16 | $19 - 3 + 0 - 0$ | $20 - 4 - 0$ |
| 4 | 10 | 2 | 15 | $16 - 1 + 0 - 0$ | $20 - 5 - 0$ |
| 5 | 12 | 2 | 13 | $15 - 2 + 0 - 0$ | $20 - 7 - 0$ |
| 6 | 13 | 1 | 10 | $13 - 2 + 0 - 1$ | $20 - 9 - 1$ |
| 7 | 14 | 1 | 9 | $10 - 1 + 0 - 0$ | $20 - 10 - 1$ |
| 8 | 15 | 1 | 8 | $9 - 1 + 0 - 0$ | $20 - 11 - 1$ |
| 9 | 18 | 1 | 4 | $8 - 1 + 0 - 3$ | $20 - 12 - 4$ |

### 10.2.4 Grouped Data

- Sometimes we work with grouped observations rather than individual observations.

- Let the values of the failure-time or loss data be divided into $k$ intervals: $(c_0, c_1]$, $(c_1, c_2], \cdots, (c_{k-1}, c_k]$, where $0 \leq c_0 < c_1 < \cdots < c_k$.

- The observations are classified into the interval groups according to the values of $x_i$ (failure time or loss).

- We first consider complete data. Let there be $n$ observations of $x_i$ in the sample, with $n_j$ observations of $x_i$ in interval $(c_{j-1}, c_j]$, so that $\sum_{j=1}^{k} n_j = n$.

- The risk set in interval $(c_0, c_1]$ is $n$. The risk set in interval $(c_1, c_2]$ is $n - n_1$.

- In general, the risk set in interval $(c_{j-1}, c_j]$ is $n - \sum_{i=1}^{j-1} n_i = \sum_{i=j}^{k} n_i$.

- When the data are incomplete, with possible left truncation and/or right censoring, approximations may be required to compute the risk sets.

- We first define the following quantities based on the attributes of individual observations

$D_j$ = number of observations with $c_{j-1} \le d_i < c_j$, for $j = 1, \cdots, k$.
$U_j$ = number of observations with $c_{j-1} < u_i \le c_j$, for $j = 1, \cdots, k$.
$V_j$ = number of observations with $c_{j-1} < x_i \le c_j$, for $j = 1, \cdots, k$.

- Thus, $D_j$ is the number of new additions to the risk set in the interval $(c_{j-1}, c_j]$, $U_j$ is the number of right-censored observations that exit the sample in the interval $(c_{j-1}, c_j]$, and $V_j$ is the number of deaths or loss values in $(c_{j-1}, c_j]$.

- We now define $R_j$ as the risk set for the interval $(c_{j-1}, c_j]$, which is the total number of observations in the sample exposed to the risk of failure or loss in $(c_{j-1}, c_j]$.

- For the first interval $(c_0, c_1]$, we have $R_1 = D_1$. Subsequent updating of the risk set is computed as

$$R_j = R_{j-1} - V_{j-1} + D_j - U_{j-1}, \qquad j = 2, \cdots, k. \tag{10.10}$$

An alternative formula for $R_j$ is

$$R_j = \sum_{i=1}^{j} D_i - \sum_{i=1}^{j-1} (V_i + U_i), \qquad j = 2, \cdots, k. \tag{10.11}$$

**Example 10.9:** For the data in Table 10.6, the observations are grouped into the intervals: $(0, 4]$, $(4, 8]$, $(8, 12]$, $(12, 16]$ and $(16, 20]$. Determine the risk set in each interval.

**Solution:** We tabulate the results as follows

**Table 10.9:** Results of Example 10.9

| Group $j$ | $D_j$ | $U_j$ | $V_j$ | $R_j$ |
|---|---|---|---|---|
| $(0, 4]$ | 13 | 0 | 0 | 13 |
| $(4, 8]$ | 7 | 0 | 4 | 20 |
| $(8, 12]$ | 0 | 1 | 5 | 16 |
| $(12, 16]$ | 0 | 3 | 3 | 10 |
| $(16, 20]$ | 0 | 3 | 1 | 4 |

$D_j$ and $U_j$ are obtained from the $d$ and $u$ values in Table 10.6, respectively. Likewise, $V_j$ are obtained by accumulating the $w$ values in Table 10.7. □