Nonlife Actuarial Models

Chapter 11

Nonparametric Model Estimation

Learning Objectives

1. Empirical distribution

2. Moments and df of the empirical distribution

3. Kernel estimates of df and pdf

4. Kaplan-Meier (product-limit) estimator and Nelson-Aalen estimator

5. Greenwood formula

6. Estimation based on grouped observations

11.1 Estimation with Complete Individual Data

11.1.1 Empirical Distribution

- We have a sample of n observations of failure times or losses X, denoted by $x_1, \cdots, x_n$.

- The distinct values of the observations are arranged in increasing order and are denoted by $0 < y_1 < \cdots < y_m$, where $m \leq n$. The value of y_j is repeated w_j times, so that $\sum_{j=1}^{m} w_j = n$.

- We also denote g_j as the partial sum of the number of observations not more than y_j, i.e., $g_j = \sum_{h=1}^{j} w_h$.

- The **empirical distribution** of the data is defined as the discrete distribution which can take values $y_1, \cdots, y_m$ with probabilities $w_1/n, \cdots, w_m/n$, respectively.

- Also, it is a discrete distribution for which the values $x_1, \cdots, x_n$ (with possible repetitions) occur with equal probabilities.

- Denoting $\hat{f}(\cdot)$ as the **empirical pf** and $\hat{F}(\cdot)$ as the **empirical df**, respectively, these functions are given by

$$\hat{f}(y) = \begin{cases} \dfrac{w_j}{n}, & \text{if } y = y_j \text{ for some } j, \\ 0, & \text{otherwise,} \end{cases} \tag{11.1}$$

and

$$\hat{F}(y) = \begin{cases} 0, & \text{for } y < y_1, \\ \dfrac{g_j}{n}, & \text{for } y_j \leq y < y_{j+1},\ j = 1, \cdots, m-1, \\ 1, & \text{for } y_m \leq y. \end{cases} \tag{11.2}$$

- Thus, the **mean of the empirical distribution** is

$$\sum_{j=1}^{m} \frac{w_j}{n} y_j = \frac{1}{n} \sum_{i=1}^{n} x_i, \tag{11.3}$$

which is the sample mean of $x_1, \cdots, x_n$, i.e., $\bar{x}$.

- The **variance of the empirical distribution** is

$$\sum_{j=1}^{m} \frac{w_j}{n} (y_j - \bar{x})^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2, \tag{11.4}$$

which is not equal to the sample variance of $x_1, \cdots, x_n$, and is biased for the variance of X.

- Estimates of the moments of X can be computed from their sample analogues. In particular, censored moments can be estimated from the censored sample.

- For example, for a policy with policy limit u, the censored kth moment $\mathrm{E}\left[(X \wedge u)^k\right]$ can be estimated by

$$\sum_{j=1}^{r} \frac{w_j}{n} y_j^k + \frac{n - g_r}{n} u^k, \qquad \text{where } y_r \leq u < y_{r+1} \text{ for some } r. \quad (11.5)$$

- The **empirical survival function** of X is $\hat{S}(y) = 1 - \hat{F}(y)$, which is an estimate of $\Pr(X > y)$.

- To compute an estimate of the df for a value of y not in the set $y_1, \cdots, y_m$, we may *smooth* the empirical df to obtain $\tilde{F}(y)$ as follows

$$\tilde{F}(y) = \frac{y - y_j}{y_{j+1} - y_j} \hat{F}(y_{j+1}) + \frac{y_{j+1} - y}{y_{j+1} - y_j} \hat{F}(y_j), \qquad (11.6)$$

where $y_j \leq y < y_{j+1}$ for some $j = 1, \cdots, m - 1$.

- $\tilde{F}(y)$ is the linear interpolation of $\hat{F}(y_{j+1})$ and $\hat{F}(y_j)$, called the **smoothed empirical distribution function**.

- To estimate the quantiles of the distribution, we also use interpolation.

- Recall that the quantile x_δ is defined as $F^{-1}(\delta)$. We use y_j as an estimate of the $(g_j/(n+1))$-quantile (or the $(100g_j/(n+1))$th percentile) of X.

- The δ-quantile of X, denoted by $\hat{x}_\delta$, may be computed as

$$\hat{x}_\delta = \left[\frac{(n+1)\delta - g_j}{w_{j+1}}\right] y_{j+1} + \left[\frac{g_{j+1} - (n+1)\delta}{w_{j+1}}\right] y_j, \tag{11.7}$$

where

$$\frac{g_j}{n+1} \leq \delta < \frac{g_{j+1}}{n+1}, \qquad \text{for some } j. \tag{11.8}$$

- Thus, $\hat{x}_\delta$ is a smoothed estimate of the sample quantiles, and is obtained by linearly interpolating y_j and y_{j+1}.

- When there are no ties in the observations, $w_j = 1$ and $g_j = j$ for $j = 1, \cdots, n$. Equation (11.7) then reduces to

$$\hat{x}_\delta = [(n+1)\delta - j]\, y_{j+1} + [j + 1 - (n+1)\delta]\, y_j, \tag{11.9}$$

where

$$\frac{j}{n+1} \leq \delta < \frac{j+1}{n+1}, \qquad \text{for some } j. \tag{11.10}$$

Example 11.1: A sample of losses has the following 10 observations

2, 4, 5, 8, 8, 9, 11, 12, 12, 16.

Plot the empirical distribution function, the smoothed empirical distribution function and the smoothed quantile function. Determine the estimates $\tilde{F}(7.2)$ and $\hat{x}_{0.75}$. Also, estimate the censored variance $\mathrm{Var}[(X \wedge 11.5)]$.

Solution: The plots of various functions are given in Figure 11.1. The empirical distribution function is a step function represented by the solid lines. The dashed line represents the smoothed empirical df, and the dotted line gives the (inverse) of the quantile function.For $\tilde{F}(7.2)$, we first note that $\hat{F}(5) = 0.3$ and $\hat{F}(8) = 0.5$. Thus, using equation (11.6) we have

$$\begin{aligned}
\tilde{F}(7.2) &= \left[\frac{7.2-5}{8-5}\right]\hat{F}(8) + \left[\frac{8-7.2}{8-5}\right]\hat{F}(5) \\
&= \left[\frac{2.2}{3}\right](0.5) + \left[\frac{0.8}{3}\right](0.3) \\
&= 0.4467.
\end{aligned}$$

For $\hat{x}_{0.75}$, we first note that $g_6 = 7$ and $g_7 = 9$ (note that $y_6 = 11$ and $y_7 = 12$). With $n = 10$, we have

$$\frac{7}{11} \leq 0.75 < \frac{9}{11},$$

so that j defined in equation (11.8) is 6. Hence, using equation (11.7), we compute the smoothed quantile as

$$\hat{x}_{0.75} = \left[\frac{(11)(0.75) - 7}{2}\right](12) + \left[\frac{9 - (11)(0.75)}{2}\right](11) = 11.625.$$

We estimate the first moment of the censored loss $\mathrm{E}\left[(X \wedge 11.5)\right]$ by

$$(0.1)(2)+(0.1)(4)+(0.1)(5)+(0.2)(8)+(0.1)(9)+(0.1)(11)+(0.3)(11.5) = 8.15,$$

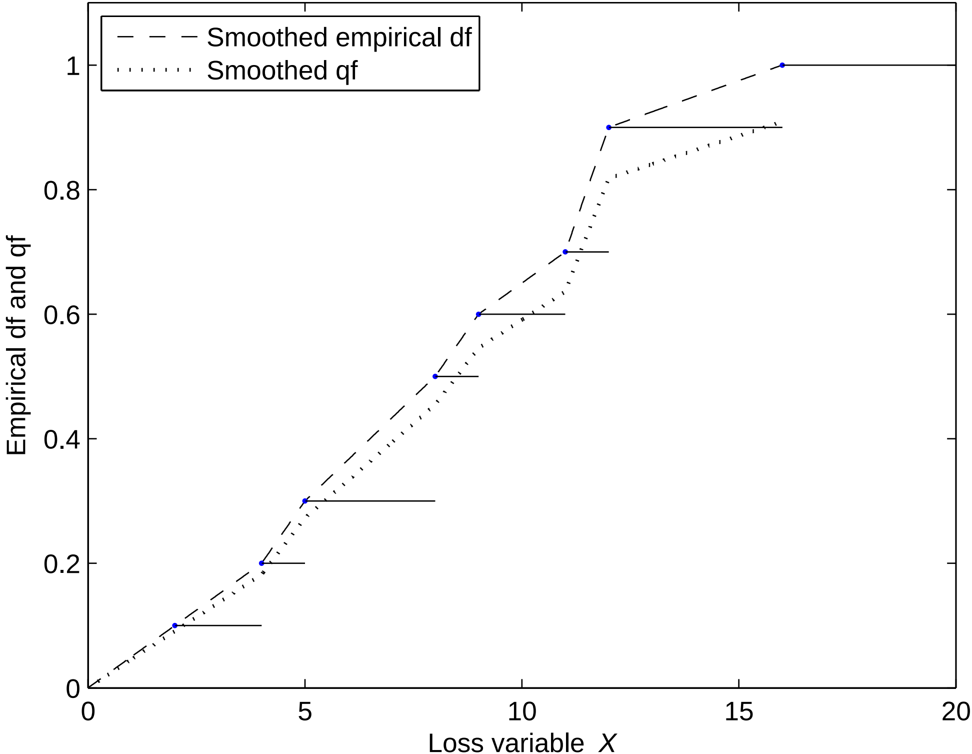
and the second raw moment of the censored loss $\mathrm{E}\left[(X \wedge 11.5)^2\right]$ by

$$(0.1)(2)^2+(0.1)(4)^2+(0.1)(5)^2+(0.2)(8)^2+(0.1)(9)^2+(0.1)(11)^2+(0.3)(11.5)^2 = 77.175.$$

Hence, the estimated variance of the censored loss is

$$77.175 - (8.15)^2 = 10.7525.$$

$\square$



Smoothed empirical df
Smoothed qf
Empirical df and qf
Loss variable X
0
0.2
0.4
0.6
0.8
1
0
5
10
15
20

- In large samples an approximate $100(1-\alpha)\%$ confidence interval estimate of $F(y)$ may be computed as

$$\hat{F}(y) \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{F}(y)[1-\hat{F}(y)]}{n}}. \tag{11.14}$$

- A drawback of (11.14) is that it may fall outside the interval $(0,1)$.

11.1.2 Kernel Estimation of Probability Density Function

- The empirical pf summarizes the data as a discrete distribution.

- If the variable of interest (loss or failure time) is continuous, it is desirable to estimate a pdf. This can be done using the **kernel density estimation method**.

- Consider the observation x_i in the sample. The empirical pf assigns a probability mass of $1/n$ to the point x_i. Given that X is continuous, we may wish to *distribute* the probability mass to a neighborhood of x_i rather than assigning it completely to point x_i.

- Let us assume that we wish to distribute the mass *evenly* in the interval $[x_i - b, x_i + b]$ for a given value of b, called the **bandwidth**. To do this, we define a function $f_i(x)$ as follows

$$f_i(x) = \begin{cases} \dfrac{0.5}{b}, & \text{for } x_i - b \leq x \leq x_i + b, \\ 0, & \text{otherwise.} \end{cases} \tag{11.15}$$

- This function is rectangular in shape, with a base of length $2b$ and height of $0.5/b$, so that its area is 1.

- It may be interpreted as the pdf contributed by the observation x_i.

- Note that $f_i(x)$ is also the pdf of a $\mathcal{U}(x_i - b, x_i + b)$ variable. Thus, only values of x in the interval $[x_i - b, x_i + b]$ receive contributions from x_i.

- As each x_i contributes a probability mass of $1/n$, the pdf of X may be estimated as

$$\tilde{f}(x) = \frac{1}{n} \sum_{i=1}^{n} f_i(x). \tag{11.16}$$

- We now rewrite $f_i(x)$ in equation (11.15) as

$$f_i(x) = \begin{cases} \dfrac{0.5}{b}, & \text{for } -1 \leq \dfrac{x - x_i}{b} \leq 1, \\ 0, & \text{otherwise,} \end{cases} \tag{11.17}$$

and define

$$K_R(\psi) = \begin{cases} 0.5, & \text{for } -1 \leq \psi \leq 1, \\ 0, & \text{otherwise.} \end{cases} \tag{11.18}$$

- Then it can be seen that

$$f_i(x) = \frac{1}{b} K_R(\psi_i), \tag{11.19}$$

where

$$\psi_i = \frac{x - x_i}{b}. \tag{11.20}$$

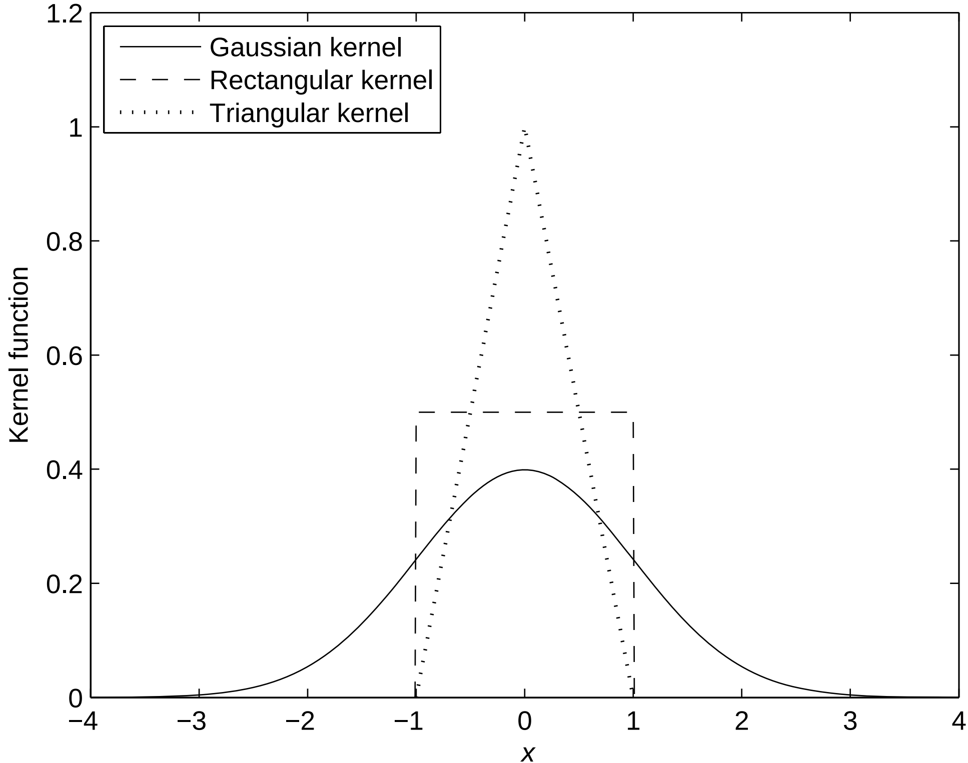
- Using equation (11.19), we rewrite equation (11.16) as

$$\tilde{f}(x) = \frac{1}{nb} \sum_{i=1}^{n} K_R(\psi_i). \tag{11.21}$$

- $K_R(\psi)$ as defined in equation (11.18) is called the **rectangular** (or **box**, **uniform**) **kernel function**. $\tilde{f}(x)$ defined in equation (11.21) is the estimate of the pdf of X using the rectangular kernel.

- It can be seen that $K_R(\psi)$ satisfies the following properties

$$K_R(\psi) \geq 0, \qquad \text{for} - \infty < \psi < \infty, \tag{11.22}$$



Gaussian kernel
Rectangular kernel
Triangular kernel
Kernel function
x
1.2
1
0.8
0.6
0.4
0.2
0
−4
−3
−2
−1
0
1
2
3
4

and

$$\int_{-\infty}^{\infty} K_R(\psi)\, d\psi = 1. \tag{11.23}$$

- Hence, $K_R(\psi)$ is itself the pdf of a random variable taking values over the real line.

- Any function $K(\psi)$ satisfying equations (11.22) and (11.23) may be called a **kernel function**.

- The expression in equation (11.21), with $K(\psi)$ replacing $K_R(\psi)$ and ψ_i defined in equation (11.20), is called the **kernel estimate** of the pdf.

- Apart from the rectangular kernel, two other commonly used kernels are the **triangular kernel**, denoted by $K_T(\psi)$, and the **Gaussian**

kernel, denoted by $K_G(\psi)$. The triangular kernel is defined as

$$K_T(\psi) = \begin{cases} 1 - |\psi|, & \text{for } -1 \leq \psi \leq 1, \\ 0, & \text{otherwise,} \end{cases} \tag{11.24}$$

and the Gaussian kernel is given by

$$K_G(\psi) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\psi^2}{2}\right), \qquad \text{for } -\infty < \psi < \infty, \tag{11.25}$$

which is just the standard normal density function.

- Figure 11.2 presents the plots of the rectangular, triangular and Gaussian kernels.

Example 11.2: A sample of losses has the following 10 observations

5, 6, 6, 7, 8, 8, 10, 12, 13, 15.

Determine the kernel estimate of the pdf of the losses using the rectangular kernel for $x = 8.5$ and 11.5 with a bandwidth of 3.

Solution: For $x = 8.5$ with $b = 3$, there are 6 observations within the interval $[5.5, 11.5]$. From equation (11.21) we have

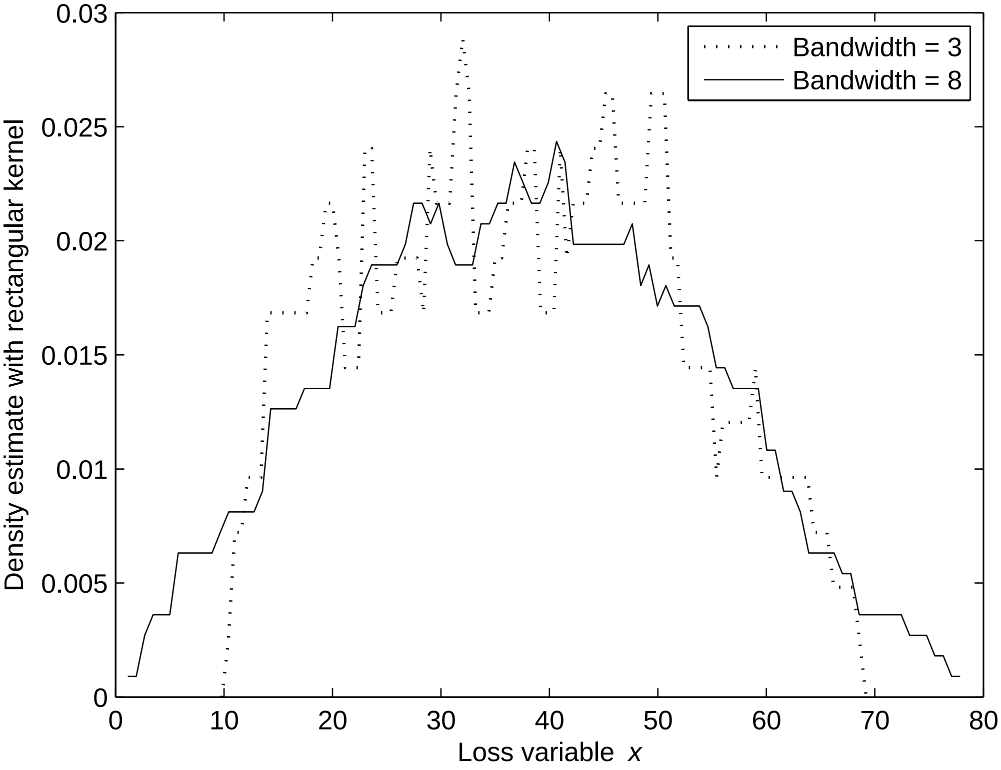
$$\tilde{f}(8.5) = \frac{1}{(10)(3)}(6)(0.5) = \frac{1}{10}.$$

Similarly, there are 3 observations in the interval $[8.5, 14.5]$, so that

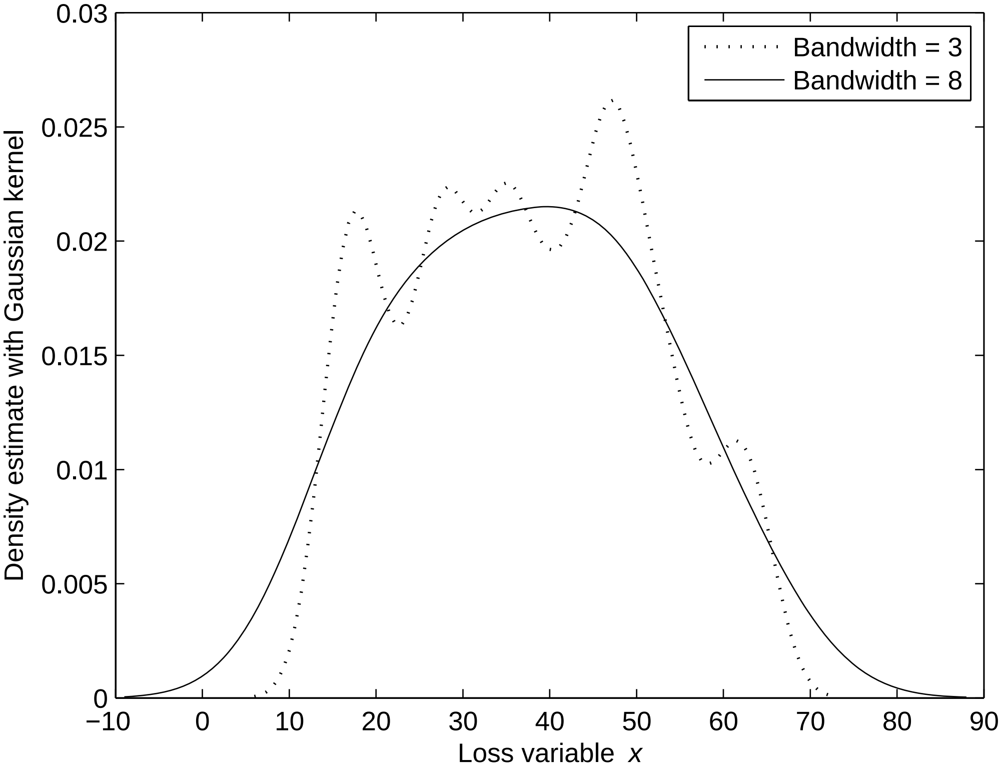
$$\tilde{f}(11.5) = \frac{1}{(10)(3)}(3)(0.5) = \frac{1}{20}.$$

□

Figures 11.3 and 11.4 show the kernel estimates of a sample of 40 observations in Example 11.3.



Bandwidth = 3
Bandwidth = 8
Density estimate with rectangular kernel
Loss variable x
0.03
0.025
0.02
0.015
0.01
0.005
0
0
10
20
30
40
50
60
70
80



Bandwidth = 3
Bandwidth = 8
Density estimate with Gaussian kernel
Loss variable x
0.03
0.025
0.02
0.015
0.01
0.005
0
−10
0
10
20
30
40
50
60
70
80
90

11.2 Estimation with Incomplete Individual Data

11.2.1 Kaplan-Meier (Product-Limit) Estimator

- We consider the estimation of $S(y_j) = \Pr(X > y_j)$, for $j = 1, \cdots, m$.
- Using the rule of conditional probability, we have

$$\begin{aligned} S(y_j) &= \Pr(X > y_1) \Pr(X > y_2 \,|\, X > y_1) \cdots \Pr(X > y_j \,|\, X > y_{j-1}) \\ &= \Pr(X > y_1) \prod_{h=2}^{j} \Pr(X > y_h \,|\, X > y_{h-1}). \end{aligned} \tag{11.27}$$

- As the risk set for y_1 is r_1 and w_1 observations are found to have value y_1, $\Pr(X > y_1)$ can be estimated by

$$\widehat{\Pr}(X > y_1) = 1 - \frac{w_1}{r_1}. \tag{11.28}$$

- Likewise, $\Pr(X > y_h \,|\, X > y_{h-1})$ can be estimated by

$$\widehat{\Pr}(X > y_h \,|\, X > y_{h-1}) = 1 - \frac{w_h}{r_h}, \qquad \text{for } h = 2, \cdots, m. \tag{11.29}$$

- Hence, we may estimate $S(y_j)$ by

$$\begin{aligned} \hat{S}(y_j) &= \widehat{\Pr}(X > y_1) \prod_{h=2}^{j} \widehat{\Pr}(X > y_h \,|\, X > y_{h-1}) \\ &= \prod_{h=1}^{j} \left(1 - \frac{w_h}{r_h}\right). \end{aligned} \tag{11.30}$$

- We now summarize the above arguments and define the Kaplan-

Meier estimator, denoted by $\hat{S}_K(y)$, as follows

$$\hat{S}_K(y) = \begin{cases} 1, & \text{for } 0 < y < y_1, \\ \prod_{h=1}^{j} \left(1 - \dfrac{w_h}{r_h}\right), & \text{for } y_j \le y < y_{j+1},\ j = 1, \cdots, m-1, \\ \prod_{h=1}^{m} \left(1 - \dfrac{w_h}{r_h}\right), & \text{for } y_m \le y. \end{cases} \tag{11.31}$$

- Note that if $w_m = r_m$, then $\hat{S}_K(y) = 0$ for $y_m \le y$.
- If $w_m < r_m$ (i.e., the largest observation is a censored observation and not a failure time), then $\hat{S}_K(y_m) > 0$. We may adopt the definition in equation (11.31).or let $\hat{S}_K(y) = 0$ for $y > y_m$, or allow $\hat{S}_K(y)$ to decay geometrically to 0 by defining

$$\hat{S}_K(y) = \hat{S}_K(y_m)^{\frac{y}{y_m}}, \qquad \text{for } y > y_m. \tag{11.32}$$

Example 11.5: Refer to the loss claims in Example 10.8. Determine the Kaplan-Meier estimate of the sf.

Solution: As all policies are with a deductible of 4, we can only estimate the conditional sf $S(y \,|\, y > 4)$. Also, as there is a maximum covered loss of 20 for all policies, we can only estimate the conditional sf up to $S(20 \,|\, y > 4)$. Using the data compiled in Table 10.8, the Kaplan-Meier estimates are summarized in Table 11.2.

Table 11.2: Kaplan-Meier estimates of Example 11.5

Interval containing y	$\hat{S}_K(y \mid y > 4)$
$(4, 5)$	1
$[5, 7)$	0.9333
$[7, 8)$	0.8667
$[8, 10)$	0.8000
$[10, 16)$	0.6667
$[16, 17)$	0.6000
$[17, 19)$	0.4000
$[19, 20)$	0.3333
20	0.2667

- The variance estimate of the Kaplan-Meier estimator can be computed as

$$\widehat{\mathrm{Var}}\left[\hat{S}_K(y_j)\,|\,\mathcal{C}\right] \simeq [\hat{S}_K(y_j)]^2 \left(\sum_{h=1}^{j} \frac{w_h}{r_h(r_h - w_h)}\right), \qquad (11.45)$$

(see NAM for the proof) which is called the **Greenwood approximation** for the variance of the Kaplan-Meier estimator.

Example 11.6: Refer to the loss claims in Examples 10.7 and 11.4. Determine the approximate variance of $\hat{S}_K(10.5)$ and the 95% confidence interval of $S_K(10.5)$.

Solution: From Table 11.1, we can see that Kaplan-Meier estimate of $S_K(10.5)$ is 0.65. The Greenwood approximate for the variance of $\hat{S}_K(10.5)$ is

$$(0.65)^2\left[\frac{1}{(20)(19)} + \frac{3}{(19)(16)} + \frac{1}{(16)(15)} + \frac{2}{(15)(13)}\right] = 0.0114.$$

Thus, the estimate of the standard deviation of $\hat{S}_K(10.5)$ is $\sqrt{0.0114} = 0.1067$, and, assuming the normality of $\hat{S}_K(10.5)$, the 95% confidence interval of $S_K(10.5)$ is

$$0.65 \pm (1.96)(0.1067) = (0.4410,\ 0.8590).$$

$\square$

- The above example uses the normal approximation for the distribution of $\hat{S}_K(y_j)$ to compute the confidence interval of $S(y_j)$. This is sometimes called the **linear confidence interval**.

- A disadvantage of this estimate is that the computed confidence interval may fall outside the range $(0, 1)$.

- This drawback can be remedied by considering a transformation of the survival function. We first define the transformation $\zeta(\cdot)$ by

$$\zeta(x) = \log\left[-\log(x)\right], \tag{11.47}$$

and let

$$\hat{\zeta} = \zeta(\hat{S}(y)) = \log[-\log(\hat{S}(y))], \tag{11.48}$$

where $\hat{S}(y)$ is an estimate of the sf $S(y)$ for a given y.

- A $100(1-\alpha)\%$ confidence interval of $S(y)$ can be computed as

$$\left(\hat{S}(y)^U, \, \hat{S}(y)^{\frac{1}{U}}\right), \tag{11.56}$$

where

$$U = \exp\left[z_{1-\frac{\alpha}{2}}\sqrt{\hat{V}(y)}\right], \tag{11.57}$$

(see NAM for a proof). This is known as the **logarithmic transformation method**.

Example 11.7: Refer to the loss claims in Examples 10.8 and 11.5. Determine the approximate variance of $\hat{S}_K(7)$ and the 95% confidence interval of $S(7)$.

Solution: From Table 11.2, we have $\hat{S}_K(7) = 0.8667$. The Greenwood approximate variance of $\hat{S}_K(7)$ is

$$(0.8667)^2 \left[\frac{1}{(15)(14)} + \frac{1}{(14)(13)}\right] = 0.0077.$$

Using normal approximation to the distribution of $\hat{S}_K(7)$, the 95% confidence interval of $S(7)$ is

$$0.8667 \pm 1.96\sqrt{0.0077} = (0.6947,\ 1.0387).$$

Thus, the upper limit exceeds 1, which is undesirable. To apply the logarithmic transformation method, we compute $\hat{V}(7)$ in equation (11.52) toobtain

$$\hat{V}(7) = \frac{0.0077}{[0.8667\,(\log 0.8667)]^2} = 0.5011,$$

so that U in equation (11.57) is

$$\exp\left[(1.96)\sqrt{0.5011}\right] = 4.0048.$$

From (11.56), the 95% confidence interval of $S(7)$ is

$$\{(0.8667)^{4.0048}, (0.8667)^{\frac{1}{4.0048}}\} = (0.5639,\ 0.9649),$$

which is within the range $(0, 1)$.

We finally remark that as all policies in this example have a deductible of 4. The sf of interest is conditional on the loss exceeding 4. $\square$

11.2.2 Nelson-Aalen Estimator

- The cumulative hazard function $H(y)$ is

$$H(y) = \int_0^y h(y)\, dy, \tag{11.58}$$

so that

$$S(y) = \exp\left[-H(y)\right]. \tag{11.59}$$

and

$$H(y) = -\log\left[S(y)\right]. \tag{11.60}$$

- If we use $\hat{S}_K(y)$ to estimate $S(y)$ for $y_j \leq y < y_{j+1}$, an estimate of the cumulative hazard function can be computed as

$$\begin{aligned}
\hat{H}(y) &= -\log\left[\hat{S}_K(y)\right] \\
&= -\log\left[\prod_{h=1}^{j}\left(1-\frac{w_h}{r_h}\right)\right] \\
&= -\sum_{h=1}^{j}\log\left(1-\frac{w_h}{r_h}\right).
\end{aligned} \tag{11.61}$$

- Using the approximation

$$-\log\left(1-\frac{w_h}{r_h}\right) \simeq \frac{w_h}{r_h}, \tag{11.62}$$

we obtain $\hat{H}(y)$ as

$$\hat{H}(y) = \sum_{h=1}^{j}\frac{w_h}{r_h}, \tag{11.63}$$

which is the **Nelson-Aalen estimate of the cumulative hazard function**.

- We complete its formula as follows:

$$\hat{H}(y) = \begin{cases} 0, & \text{for } 0 < y < y_1, \\ \sum_{h=1}^{j} \dfrac{w_h}{r_h}, & \text{for } y_j \leq y < y_{j+1},\ j = 1, \cdots, m-1, \\ \sum_{h=1}^{m} \dfrac{w_h}{r_h}, & \text{for } y_m \leq y. \end{cases} \tag{11.64}$$

- The **Nelson-Aalen estimator of the survival function**, denoted by $\hat{S}_N(y)$ is

$$\hat{S}_N(y) = \begin{cases} 1, & \text{for } 0 < y < y_1, \\ \exp\left(-\sum_{h=1}^{j} \dfrac{w_h}{r_h}\right), & \text{for } y_j \le y < y_{j+1},\ j = 1, \cdots, m-1, \\ \exp\left(-\sum_{h=1}^{m} \dfrac{w_h}{r_h}\right), & \text{for } y_m \le y. \end{cases} \tag{11.65}$$

- For $y > y_m$, we may also compute $\hat{S}_N(y)$ as 0 or $[\hat{S}_N(y_m)]^{\frac{y}{y_m}}$.

- In the case of complete data, with one observation at each point y_j, we have $w_h = 1$ and $r_h = n - h + 1$ for $h = 1, \cdots, n$, so that

$$\hat{S}_N(y_j) = \exp\left(-\sum_{h=1}^{j} \frac{1}{n-h+1}\right). \tag{11.66}$$

- To derive an approximate formula for the variance of $\hat{H}(y)$, we assume the conditional distribution of W_h given the information set $\mathcal{C}$

to be Poisson.

- We estimate $\text{Var}(W_h)$ by w_h. An estimate of $\text{Var}[\hat{H}(y_j)]$ can then be computed as

$$\widehat{\text{Var}}[\hat{H}(y_j)] = \widehat{\text{Var}}\left(\sum_{h=1}^{j} \frac{W_h}{r_h}\right) = \sum_{h=1}^{j} \frac{\widehat{\text{Var}}(W_h)}{r_h^2} = \sum_{h=1}^{j} \frac{w_h}{r_h^2}. \qquad (11.67)$$

- A100$(1-\alpha)$% confidence interval of $H(y_j)$, assuming normal approximation, is given by

$$\hat{H}(y_j) \pm z_{1-\frac{\alpha}{2}} \sqrt{\widehat{\text{Var}}[\hat{H}(y_j)]}. \qquad (11.68)$$

- To ensure the lower limit of the confidence interval of $H(y_j)$ to be positive, we consider the transformation

$$\zeta(x) = \log(x), \qquad (11.69)$$

and a $100(1-\alpha)\%$ approximate confidence interval of $H(y_j)$ is

$$\left(\hat{H}(y_j)\left(\frac{1}{U}\right), \hat{H}(y_j)U\right), \tag{11.73}$$

where

$$U = \exp\left[z_{1-\frac{\alpha}{2}} \frac{\sqrt{\widehat{\mathrm{Var}}[\hat{H}(y_j)]}}{\hat{H}(y_j)}\right]. \tag{11.74}$$

11.3 Estimation with Grouped Data

- We assume that the values of the failure-time or loss data x_i are grouped into k intervals: $(c_0, c_1]$, $(c_1, c_2], \cdots, (c_{k-1}, c_k]$, where $0 \leq c_0 < c_1 < \cdots < c_k$.

- We first consider the case where the data are complete, with no truncation nor censoring.

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

- Assuming the observations within each interval are uniformly distributed, the empirical pdf of the failure-time or loss variable X can be written as

$$\hat{f}(x) = \sum_{j=1}^{k} p_j f_j(x), \tag{11.75}$$

where

$$p_j = \frac{n_j}{n} \tag{11.76}$$

and

$$f_j(x) = \begin{cases} \dfrac{1}{c_j - c_{j-1}}, & \text{for } c_{j-1} < x \le c_j, \\ 0, & \text{otherwise.} \end{cases} \tag{11.77}$$

- Thus, $\hat{f}(x)$ is the pdf of a mixture distribution. To compute the moments of X we note that

$$\int_0^{\infty} f_j(x) x^r \, dx = \frac{1}{c_j - c_{j-1}} \int_{c_{j-1}}^{c_j} x^r \, dx = \frac{c_j^{r+1} - c_{j-1}^{r+1}}{(r+1)(c_j - c_{j-1})}. \tag{11.78}$$

- Hence, the mean of the empirical pdf is

$$\mathrm{E}(X) = \sum_{j=1}^{k} p_j \left[\frac{c_j^2 - c_{j-1}^2}{2(c_j - c_{j-1})}\right] = \sum_{j=1}^{k} \frac{n_j}{n}\left[\frac{c_j + c_{j-1}}{2}\right], \tag{11.79}$$

and its rth raw moment is

$$\mathrm{E}(X^r) = \sum_{j=1}^{k} \frac{n_j}{n}\left[\frac{c_j^{r+1} - c_{j-1}^{r+1}}{(r+1)(c_j - c_{j-1})}\right]. \tag{11.80}$$

- The censored moments are more complex. Suppose it is desired to compute $\mathrm{E}[(X \wedge u)^r]$. First, we consider the case where $u = c_h$ for some $h = 1, \cdots, k-1$, i.e., u is the end point of an interval.

- Then the rth raw moment is

$$E\left[(X \wedge c_h)^r\right] = \sum_{j=1}^{h} \frac{n_j}{n}\left[\frac{c_j^{r+1} - c_{j-1}^{r+1}}{(r+1)(c_j - c_{j-1})}\right] + c_h^r \sum_{j=h+1}^{k} \frac{n_j}{n}. \tag{11.81}$$

- If $c_{h-1} < u < c_h$, for some $h = 1, \cdots, k$, then we have

$$E\left[(X \wedge u)^r\right] = \sum_{j=1}^{h-1} \frac{n_j}{n} \left[\frac{c_j^{r+1} - c_{j-1}^{r+1}}{(r+1)(c_j - c_{j-1})}\right] + u^r \sum_{j=h+1}^{k} \frac{n_j}{n} + \frac{n_h}{n(c_h - c_{h-1})} \left[\frac{u^{r+1} - c_{h-1}^{r+1}}{r+1} + u^r (c_h - u)\right]. \quad (11.82)$$

- The empirical df at the upper end of each interval is easy to compute. Specifically, we have

$$\hat{F}(c_j) = \frac{1}{n} \sum_{h=1}^{j} n_h, \qquad \text{for } j = 1, \cdots, k. \quad (11.83)$$

- For other values of x, we use the interpolation formula given in equation (11.6), i.e.,

$$\hat{F}(x) = \frac{x - c_j}{c_{j+1} - c_j}\, \hat{F}(c_{j+1}) + \frac{c_{j+1} - x}{c_{j+1} - c_j}\, \hat{F}(c_j), \quad (11.84)$$

where $c_j \leq x < c_{j+1}$, for some $j = 0, 1, \cdots, k-1$, with $\hat{F}(c_0) = 0$. $\hat{F}(x)$ is also called the **ogive**.

- When the observations are incomplete, we may use the Kaplan-Meier and Nelson-Aalen methods to estimate the sf.

- Using equations (10.10) or (10.11), we calculate the risk sets R_j and the number of failures or losses V_j in the interval $(c_{j-1}, c_j]$.

- These numbers are taken as the risk sets and observed failures or losses at points c_j. $\hat{S}_K(c_j)$ and $\hat{S}_N(c_j)$ may then be computed using equations (11.31) and (11.65), respectively, with R_h replacing r_h and V_h replacing w_h.