Assignment 1 - Time-to-Event Analysis

Arathy Bastin (03754963)
Varsha Raveendran(03745012)
Sandhanakrishnan Ravichandran (03755546)

June 7, 2022

Exercise 1

Suppose survival times of subjects are censored if they move out of the study area and thereby can no longer be followed for survival. Give examples where one could envision that such censoring is and is not informative.

Scenario: Consider a clinical trial that studies the effect of a treatment for cancer. The end point of the trial is the death of patient due to the disease. A patient who has a complete remission of cancer decides to move out of the study area to a different city.

Informative Censoring - when participants are lost to follow-up due to reasons related to the study.
Example:

- The person does not return for the follow-up though the clinical trial has follow-ups planned for 5 years. If he had returned, there was a lower chance of a death from cancer during follow-up. By censoring this person, we are likely to overestimate the mortality risk. This is informative censoring - it gives us information regarding the risk of experiencing the event (here, death caused by cancer).

Non-informative Censoring - when participants are lost to follow-up due to reasons unrelated to the study.
Example:

- If the patient dies of other causes such as a road accident, then censoring this person is non-informative. It does not give any information regarding the risk of experiencing the event/end-point.

Exercise 2

For a continuous-time random variable T, proof that 1. f(t)=−dS(t)/dt
2. H(t) = − log[S(t)]
Note that S(0)=1 and $\int_a^b f(x)\,\mathrm{d}x = \log|f(x)| + C$.

Proof:
1.

$$\mathrm{S(t)} = 1 \text{ - } \mathrm{P(T} \leq t)$$

$$S(t) = 1 - F(t) \tag{1}$$

$$S(t) = 1 - \int_0^t f(x)dx \tag{2}$$

$$\frac{dS(t)}{dt} = 0 - \frac{d}{dt}\int_0^t f(x)dx \tag{3}$$

$$\frac{dS(t)}{dt} = -f(t) \tag{4}$$

$$\boxed{f(t) = -\frac{dS(t)}{dt}} \tag{5}$$

2.

$$\text{H(t) = -log [S(t)]}$$

$$\text{The hazard function } h(t) = \frac{f(t)}{S(t)} \tag{6}$$

$$= -\frac{dS(t)}{dt}.\frac{1}{S(t)} \tag{7}$$

$$\text{We next use the chain rule of differentiation i.e } \frac{df(u)}{dx} = f'(u).\frac{d(u)}{d(x)}. \tag{8}$$

Applying differentiation over a log function we get,

$$\frac{d\log(f(x))}{dx} = \frac{d\log(f(x))}{df(x)}.\frac{df(x)}{dx} \tag{9}$$

$$\frac{d\log(f(x))}{dx} = \frac{df(x)}{dx}.\frac{1}{f(x)} \tag{10}$$

Comparing equations 7 and 4,

$$h(t) = -\frac{d\log(S(t))}{t} \tag{11}$$

Hazard rate is negative log of survival rate. Therefore, cumulative hazard rate, H(t) is the integration of h(t) over time.

Hence,

$$\boxed{H(t) = -\log[S(t)]}$$

Exercise 3

Show that $S_1(t) = [S_0(t)]^R$ is equivalent to $\frac{h_1(t)}{h_0(t)}$ = R (proportional hazards).

Proof:

The survival function is related to the cumulative hazard function H(t) by

$$\mathrm{S}_0(t) = exp(-H_0(t))$$
$$\mathrm{S}_1(t) = exp(-H_1(t))$$

$$\mathrm{S}_1(t) = [\mathrm{S}_0(t)]^R$$

$$\exp(\text{-}\mathrm{H}_1(t)) = [\exp(\text{-}\mathrm{H}_0(t))]^R$$

Taking log on both sides,

$$\log[\exp(\text{-}\mathrm{H}_1(t)) = \log[\exp(\text{-}\mathrm{H}_0(t))]^R$$

$$\text{-}\mathrm{H}_1(t) = R * [\log[\exp(\text{-}\mathrm{H}_0(t))]$$
$$\text{-}\mathrm{H}_1(t) = -R * H_0(t).$$

Differentiating H1 and H0 w.r.t to t

$$\frac{\mathrm{d}H_1(t)}{\mathrm{d}t} = \mathrm{R} * \frac{\mathrm{d}H_0(t)}{\mathrm{d}t}$$

Substituting $\mathrm{h(t)} = \frac{\mathrm{d}H(t)}{\mathrm{d}t}$ in the above equation we get,

$$\mathrm{h}_1(t) = R * h_0(t)$$

$$\frac{h_1(t)}{h_0(t)} = \mathrm{R} \text{ (proportional hazards)}$$

Exercise 4

In the hands-on exercise, we used the non-parametric Kaplan-Meier estimator to estimate survival curves from data. Here, we want to use a parametric approach based on the exponential distribution, which is defined as:
Hazard function: $h_\lambda(t) = \lambda$
Survival function: $S_\lambda(t) = exp(-\lambda t)$
Density function: $f_\lambda(t) = \lambda exp(-\lambda t)$
where $\lambda > 0$ is the unknown parameter that needs to be estimated by solving $argmax_\lambda \sum_{i=1}^{n} [\delta_i log h_\lambda(y_i) + log S_\lambda(y_i)]$.

$$Let \log L(\lambda) = \sum_{i=1}^{n} [\delta_i \log h_\lambda(y_i) + \log S_\lambda(y_i)] \tag{12}$$

$$\hat{\lambda} = argmax_\lambda \log L(\lambda)$$

To find the optimal value of λ, $\hat{\lambda}$, we differentiate equation (1) w.r.t to λ and equate it to 0. ie $\frac{\mathrm{d} log L(\lambda)}{\mathrm{d}\lambda} = 0$.
From equation (1),

$$\frac{\mathrm{d}(\sum_{i=1}^{n} [\delta_i log h_\lambda(y_i) + log S_\lambda(y_i)])}{\mathrm{d}\lambda} = 0 \tag{13}$$

$$\sum_{i=1}^{n} (\frac{\mathrm{d}([\delta_i log h_\lambda(y_i) + log S_\lambda(y_i)])}{\mathrm{d}\lambda}) = 0 \tag{14}$$

$h_\lambda(y_i) = \lambda$ and
$S_\lambda(y_i) = exp(-\lambda y_i)$
Using the above two equations in (3), we get

$$\sum_{i=1}^{n} (\frac{\mathrm{d}([\delta_i log\lambda + log(exp(-\lambda y_i))])}{\mathrm{d}\lambda}) = 0 \tag{15}$$

$$\sum_{i=1}^{n} (\frac{\mathrm{d}(\delta_i log\lambda)}{\mathrm{d}\lambda} + \frac{\mathrm{d}(log(exp(-\lambda y_i)))}{\mathrm{d}\lambda}) = 0 \tag{16}$$

$$\sum_{i=1}^{n} (\frac{\delta_i}{\lambda} - y_i) = 0 \tag{17}$$

$$\sum_{i=1}^{n} (\frac{\delta_i}{\lambda}) - \sum_{i=1}^{n} (y_i) = 0 \tag{18}$$

$$\sum_{i=1}^{n} (\frac{\delta_i}{\lambda}) = \sum_{i=1}^{n} (y_i) \tag{19}$$

$$\boxed{\hat{\lambda} = \frac{\sum_{i=1}^{n} \delta_i}{\sum_{i=1}^{n} y_i}} \tag{20}$$

Survival function:

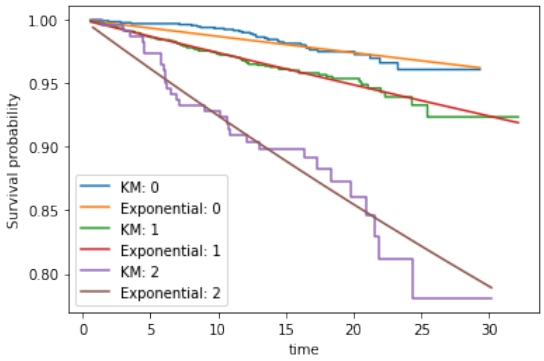
$$\boxed{S_\lambda(y_i) = exp(-\frac{\sum_{i=1}^{n} \delta_i}{\sum_{i=1}^{n} y_i} y_i)} \tag{21}$$

Exercise 5

Implement the estimator of the survival function of the exponential distribution derived in Exercise 4 above. Use the breast cancer dataset from the hands-on exercise to compare the Kaplan-Meier and parametric estimator for women with entry diagnoses of AH, PDWA, and No PD as a function of years since biopsy. What does the graphic tell? Explain whether the exponential distribution is suitable for this data?

(solution on next page)

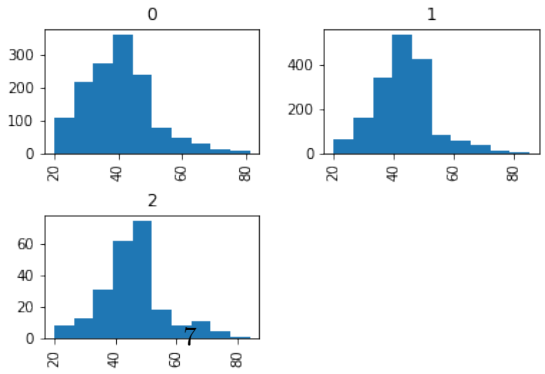
The plots for survival functions using Kaplan-Meier estimate and exponential distribution are shown below.



Survival function: Kaplan Meier estimate v/s Parametric estimate

Observations

- The exponential curves show a similar pattern to the Kaplan-Meier estimates - women with entry diagnoses of AH has a lower survival probability than PDWA and No PD. However, it is possible to predict the probability of survival from exponential curves between successive events since it is a continuous function (unlike Kaplan-Meier where it is constant between events and drops only when an event occurs).
- Here, the estimator implemented (exercise 4) considers the maximum likelihood case (optimal) assuming that risk is constant (i.e. the hazard function $h_\lambda(t) = \lambda$ does not depend on time and is constant.). The exponential curve shows a constant rate of event occurrence which is unrealistic for this data.
- The histogram shows that the groups differ in age. Due to the memoryless property of exponential distribution (i.e) the survival probability does not consider ageing of participants. It is difficult to conclude from plots of both estimators whether survival depends on only entry diagnoses or other factors such as age.



Age of participants grouped by pd

- Exponential models are useful if hazard function is constant and does not depend on time; it is not the case in our data and thus not useful.