Survival Analysis (Review & Exercise Solutions)

Alireza Ghorbani

2025-03-16

Contents

1 Introduction to Survival Analysis

1.1 What Is Survival Analysis?

Survival analysis is a collection of statistical procedures for data analysis for which the outcome variable of interest is time until an event occurs.

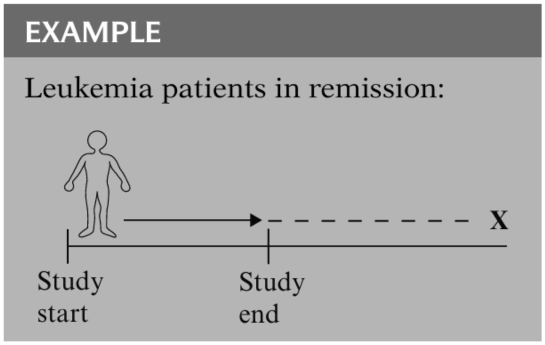
By **time**, we mean years, months, weeks, or days from the beginning of follow-up of an individual until an event occurs; alternatively, time can refer to the age of an individual when an event occurs.

By **event**, we mean death, disease incidence, relapse from remission, recovery (e.g., return to work) or any designated experience of interest that may happen to an individual.

In a survival analysis, we usually refer to the time variable as **survival time**, because it gives the time that an individual has "survived" over some follow-up period. We also typically refer to the event as a **failure**, because the event of interest usually is death, disease incidence, or some other negative individual experience. However, survival time may be "time to return to work after an elective surgical procedure," in which case failure is a positive event.

1.2 Censored Data

Most survival analyses must consider a key analytical problem called **censoring**. In essence, censoring occurs when we have some information about individual survival time, but **we don't know the survival time exactly**.

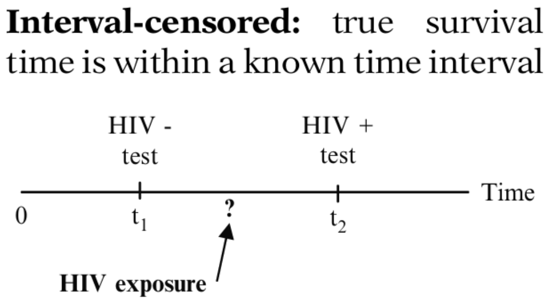


There are generally three reasons why censoring may occur: 1. A person does not experience the event before the study ends; 2. A person is lost to follow-up during the study period; 3. A person withdraws from the study because of death (if death is not the event of interest) or some other reason (e.g., adverse drug reaction or other competing risk).

We know that the person's true survival time becomes incomplete at the right side of the follow-up period, occurring when the study ends or when the person is lost to follow-up or is withdrawn. We generally refer to this kind of data as **right-censored**. For these data, the complete survival time interval, which we don't really know, has been cut off (i.e., censored) at the right side of the observed survival time interval.

Left-censored data can occur when a person's true survival time is less than or equal to that person's observed survival time. For example, if we are following persons until they become HIV positive, we may record a failure when a subject first tests positive for the virus. However, we may not know the exact time of first exposure to the virus, and therefore do not know exactly when the failure occurred. Thus, the survival time is censored on the left side since the true survival time, which ends at exposure, is shorter than the follow-up time, which ends when the subject's test is positive. In other words, if a person is left-censored at time t , we know they had an event between time 0 and t, but we do not know the exact time of the event.

Survival analysis data can also be **interval-censored**, which can occur if a subject's true (but unobserved) survival time is within a certain known specified time interval. As an example, again considering HIV surveillance, a subject may have had two HIV tests, where he/she was HIV negative at the time (say, t_1) of the first test and HIV positive at the time (t_2) of the second test. In such a case, the subject's true survival time occurred after time t_1 and before time t_2, i.e., the subject is interval-censored in the time interval (t_1, t_2).



Interval-censored: true survival time is within a known time interval

1.3 Terminology and Notation

The key notation is **T** for the survival time variable, **t** for a specified value of T, and **d** for the dichotomous variable indicating event occurrence or censorship. The key terms are the survivor function $S(t)$ and the hazard function $h(t)$, which are in essence opposed concepts, in that the survivor function focuses on surviving whereas the hazard function focuses on failing, given survival up to a certain time point.

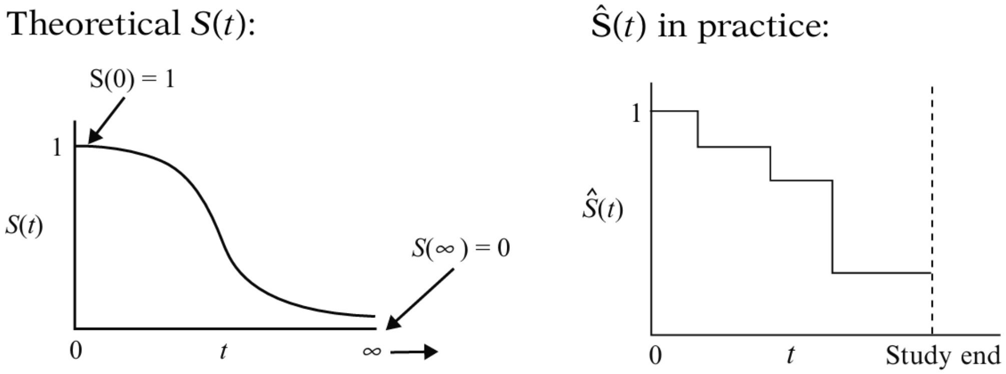
Notation	Description
T	Survival time random variable
t	Specific value of T
d	(0, 1) variable for failure/censorship
S(t)	Survivor function
h(t)	Hazard function

The **survivor function** $S(t)$ gives the probability that a person survives longer than some specified time t. That is, $S(t)$ gives the probability that the random variable T exceeds the specified time t;

$$S(t) = P(T > t).$$

Theoretically, as t ranges from 0 up to infinity, the survivor function can be graphed as a smooth curve. As illustrated by the graph, where t identifies the X-axis, all survivor functions have the following characteristics: - They are nonincreasing; that is, they head downward as t increases. - At time $t = 0$, $S(t) = S(0) = 1$; that is, at the start of the study, since no one has gotten the event yet, the probability of surviving past time 0 is one. - At time $t \to \infty$, $S(t) = 0$; that is, theoretically, if the study period increased without limit, eventually nobody would survive, so the survivor curve must eventually fall to zero.

In practice, when using actual data, we usually obtain graphs that are **step functions**, as illustrated here, rather than smooth curves. Moreover, because the study period is never infinite in length and there may be competing risks for failure, it is possible that not everyone studied gets the event. The estimated survivor function, denoted by a caret over the S in the graph, thus may not go all the way down to zero at the end of the study.



The hazard function, denoted by $h(t)$, is given by the formula:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t}$$

Because of the given formula here, the hazard function is some times called a **conditional failure rate**. The hazard function $h(t)$ gives the instantaneous potential per unit time for the event to occur, given that the individual has survived up to time t. Note that, in contrast to the survivor function, which focuses on not failing, the hazard function focuses on failing, that is, on the event occurring. Thus, in some sense, the hazard function can be considered as giving the opposite side of the information given by the survivor function.

1.4 Relationship Between Survivor Function and Hazard Function

Regardless of which function $S(t)$ or $h(t)$ one prefers, there is a clearly defined relationship between the two. In fact, if one knows the form of $S(t)$, one can derive the corresponding $h(t)$, and vice versa. For example, if the hazard function is constant, i.e., $h(t) = \lambda$ for some specific value λ, then it can be shown that the corresponding survival function is given by the following formula $S(t) = e^{-\lambda t}$

General formulae: the relationship between $S(t)$ and $h(t)$ can be expressed equivalently in either of two calculus formulae shown here. The first of these formulae describes how the survivor function $S(t)$ can be written in terms of an integral involving the hazard function:

$$S(t) = \exp\left(-\int_0^t h(u)\, du\right).$$

The second formula describes how the hazard function $h(t)$ can be written in terms of a derivative involving the survivor function:

$$h(t) = -\left(\frac{dS(t)/dt}{S(t)}\right).$$

1.5 Goals of Survival Analysis

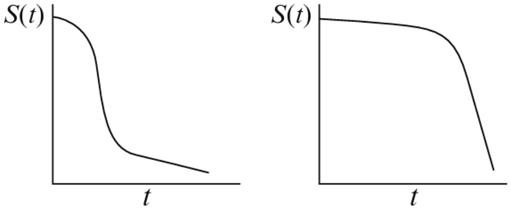
We now state the basic goals of survival analysis.

Goal 1: To estimate and interpret survivor and/or hazard functions from survival data.

Goal 2: To compare survivor and/or hazard functions.

Goal 3: To assess the relationship of explanatory variables to survival time.

Regarding the **first goal**, consider, for example, the two survivor functions pictured at the left, which give very different interpretations. The function farther on the left shows a quick drop in survival probabilities early in follow-up but a leveling off thereafter. The function on the right, in contrast, shows a very slow decrease in survival probabilities early in follow-up but a sharp decrease later on.



We compare survivor functions for a treatment group and a placebo group by graphing these functions on the same axis. Note that up to 6 weeks, the survivor function for the treatment group lies above that for the placebo group, but thereafter the two functions are at about the same level. This dual graph indicates that up to 6 weeks the treatment is more effective for survival than the placebo but has about the same effect thereafter.

Goal 3 usually requires using some form of mathematical modeling, for example, the Cox proportional hazards approach, which will be the subject of subsequent chapters.

1.6 Censoring Assumptions

Three assumptions about censoring:

1. Independent (vs. non-independent) censoring
2. Random (vs. non-random) censoring
3. Non-informative (vs. informative) censoring

Random censoring essentially means that subjects who are censored at time t should be representative of all the study subjects who remained at risk at time t with respect to their survival experience. In other words, the failure rate for subjects who are censored is assumed to be equal to the failure rate for subjects who remained in the risk set who are not censored.

Independent censoring essentially means that within any subgroup of interest, the subjects who are censored at time t should be representative of all the subjects in that subgroup who remained at risk at time t with respect to their survival experience. In other words, censoring is independent provided that it is random within any subgroup of interest.

Non-informative censoring occurs if the distribution of survival times (T) provides no information about the distribution of censorship times (C), and vice versa. Otherwise, the censoring is informative. Note, however, that the data must still identify which subjects are or are not censored.

1.7 Exercise 1

The following table shows data on lung cancer patients who underwent surgical resection as primary treatment. The outcome of interest is disease-free survival after the primary treatment, i.e. the length of time that the patient survives without any signs or symptoms of lung cancer.

Id	t_{treat}	t_{rec}	t_{death}	t_{study}	d	t
1	17	NA	NA	2098	0	2081
2	82	NA	612	612	1	530
3	51	313	811	811	1	262
4	NA	NA	29	29	NA	NA
5	84	1440	1480	1480	1	1356
6	18	1679	NA	2098	1	1661
7	21	NA	NA	356	0	335

Id: Identifier, t_{treat}: time until surgical resection, t_{rec}: time until recurrence of lung cancer, t_{death}: time until death, t_{study}: time on study

1. What is the event of interest?
2. Why are some of the values for time until recurrence and time until death missing?
3. Why is the value for time until surgical resection missing for patient 4? Is it correct to include this patient in the analyses?
4. Complete the two columns d and t in the table, which give information on whether the event of interest occurred and the time until it occurred.
5. Are the censored individuals left-, interval- or right-censored? What is the likeliest reason for censoring for each of the censored individuals?

Answer:

1. The event of interest is disease-free survival, which is defined as the time from surgical resection until the first recurrence of lung cancer or the event of death, whichever happens first.

2. Some values are missing because those events did not occur during the follow-up period. If a patient did not experience recurrence by the end of the study, t_{rec} is missing. Similarly, if they were alive at the last follow-up, t_{death} is missing. Patient 2 only died, patient 6 only experienced recurrence, and patients 1 & 7 experienced none.

3. The time until surgical resection is missing for patient 4 because they did not undergo surgery at all (or even worse missing information). Instead, they passed away very early in the study (at $t = 29$). Also, they should be excluded from the analysis because they did not receive the treatment of interest (surgical resection). Including them would bias survival estimates since their outcome is unrelated to the effect of surgery.

4. (Answers in the table): d (Event Indicator): 1 if the patient experienced recurrence, 0 if censored. t (Observed Time): The time until the event occurred or the last follow-up time for censored patients.

5. Both censored individuals, 1 & 7 are right-censored, meaning that the event of interest had not occurred by the last recorded follow-up time. The reasons for censoring could be loss to follow-up, withdraw from the study, or that study ended at some time.

2 Kaplan-Meier Survival Curves and the Log-Rank Test

2.1 General Features of KM Curves

The general formula for a KM survival probability at failure time $t(f)$ is shown here. This formula gives the probability of surviving past the previous failure time $t(f-1)$, multiplied by the conditional probability of surviving past time $t(f)$, given survival to at least time $t(f)$.

$$\hat{S}(t_f) = \hat{S}(t_{f-1}) \cdot \hat{Pr}(T > t_f | T \geq t_f)$$

The above KM formula can also be expressed as a product limit if we substitute for the survival probability $\hat{S}(t(f-1))$, the product of all fractions that estimate the conditional probabilities for failure times $t(f-1)$ and earlier.

$$\hat{S}(t_{f-1}) = \prod_{i=1}^{f} \hat{Pr}(T > t_{(i)} | T \geq t_{(i)})$$

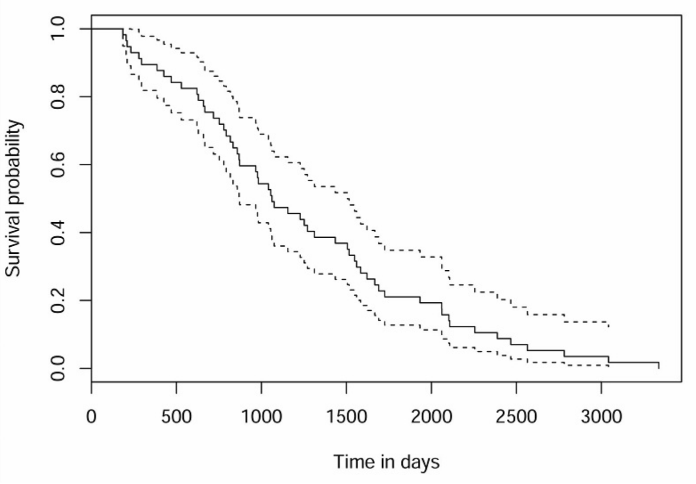
Example:

$$\hat{S}(10) = 0.8067 = \frac{14}{18} \times \frac{16}{17} \times \frac{21}{15} = 0.7529$$

$$\hat{S}(16) = 0.6902 = \frac{10}{11} = 0.8067 \times \frac{11}{12} \times \frac{10}{11}$$

2.2 Exercise 2

The following plot shows the estimated survivor curve for patients with malignant melanoma who had their tumor removed by surgery at the Department of Plastic Surgery, University Hospital of Odense, Denmark during the period 1962 to 1977. Survival time is given in days after the operation.



1. What is the median survival time?
2. What is the probability to survive at least 1500 days after the operation?
3. Out of 100 patients, how many do you expect to be dead after 24 months?
4. When does the Kaplan-Meier curve drop?
   (a) When a patient is censored

(b) When a patient dies
(c) When a patient is censored or dies
(d) None of the above

Answer:

1. The median survival time is the time at which 50% of patients are still alive and 50% have died. On the Kaplan-Meier (KM) curve where Survival Probability = 0.5 (50%) on the y-axis. From this point, we can draw a horizontal line to intersect the survival curve, then drop a vertical line to the x-axis (time in days). Based on the plot it is **almost 1000 days**.

2. On where there is 1500 days on the x-axis of the Kaplan-Meier curve, we move vertically until intersecting the curve. From this intersection, move horizontally to read the corresponding survival probability on the y-axis. Based on the plot, **the probability of surviving at least 1500 days is about 0.4**.

3. (24 months = 2 years = 730 days) Based on the Kaplan-Meier curve, the survival probability at 730 days is almost 0.75. Therefore the probability of death by 730 days is almost 0.25. And, the estimate for the number of patients expected to be dead is :

$$\text{Expected deaths} = (1 - P(\text{survival at 730 days})) \times 100 = (1 - 0.75) \times 100 = 25$$

4. The curve drops only when an event of death occurs. So **only (b)** is correct.

2.3 Exercise 3

We observe the following survival time data (in months): 12.3+, 5.4, 8.2, 12.2+, 11.7, 10+, 5.7+, 9.8, 2.6, 11.0, 9.2, 12.1+, 6.6, 2.2, 1.8, 10.2, 10.7, 11.1, 5.3, 3.5, 9.2, 2.5, 8.7, 3.8, 3+, where + denotes right-censored data.

1. Why would it be incorrect to ignore the censorship status and to calculate the average survival time by taking the average of all survival times?
2. Fill in the following table where $t_{(f)}$ are the ordered failure times, m_f are the number of failure sat time $t_{(f)}$, q_f are the number of persons who are censored in the interval $[t_{(f)}t_{(f+1)})$, $R(t_{(f)})$ is the risk set at time $t_{(f)}$ and $\hat{S}(t_{(f)})$ is the Kaplan-Meier estimate of survival at time $t_{(f)}$.

$t_{(f)}$	m_f	q_f	$\lvert R(t_{(f)})\rvert$	$\hat{S}(t_{(f)})$
1.8	1	0	25	
2.2				
2.5				
. . .				

3. Plot the corresponding Kaplan-Meier curve.

Answer:

1. Censored data means we do not observe the actual survival time for some individuals. Simply taking the average of all recorded times underestimates survival since censored observations artificially lower the calculated mean. Kaplan-Meier estimation correctly accounts for censored observations by estimating survival probabilities at different time points.
2. We do the following steps:
   (a) Remove censoring time from the list of failure times.
   (b) Order **unique** failure time from smallest to largest; count ties ony once.
   - $t_{(f)}$: ordered failure time
   - m_f: number of failure occurring at $t_{(f)}$
   - q_f: number of subjects censored in $[t_{(f)}, t_{(f+1)}]$
   - $|R(t_{(f)})|$: risk set
   - $\hat{S}(t_{(f)})$: estimated survival probability $\hat{S}(t_{(f)}) = \hat{S}(t_{(f-1)}) \times (1 - \frac{m_f}{R(t_f)})$.

   Here is the R code for the steps mentioned:

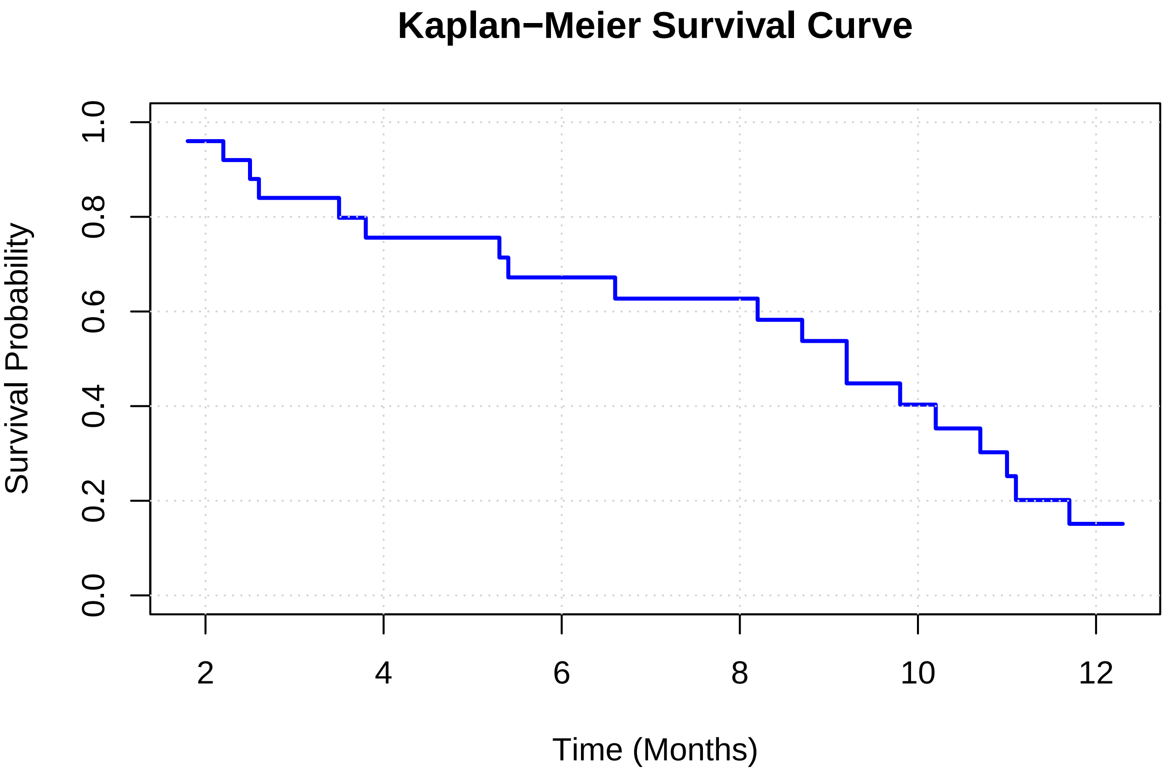
```
# Define survival times
time <- c(12.3, 5.4, 8.2, 12.2, 11.7, 10, 5.7, 9.8, 2.6, 11.0, 9.2, 12.1,
          6.6, 2.2, 1.8, 10.2, 10.7, 11.1, 5.3, 3.5, 9.2, 2.5, 8.7, 3.8, 3)
# Define censoring indicator (1 = event occurred, 0 = censored)
status <- c(0, 1, 1, 0, 1, 0, 0, 1, 1, 1, 1, 0,
            1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0)

# Combine into a dataframe
data <- data.frame(time, status)
# Sort by time
data <- data[order(data$time), ]

# Initialize variables
unique_times <- unique(data$time)  # Unique event times
n <- length(time)  # Total number of individuals
```

```
at_risk <- rep(0, length(unique_times))
events <- rep(0, length(unique_times))
censored <- rep(0, length(unique_times))
survival_prob <- rep(1, length(unique_times))

# Compute Kaplan-Meier estimates
current_surv <- 1
for (i in seq_along(unique_times)) {
  t <- unique_times[i]
  at_risk[i] <- sum(data$time >= t)
  events[i] <- sum(data$time == t & data$status == 1)
  censored[i] <- sum(data$time == t & data$status == 0)
  if (events[i] > 0) {
    current_surv <- current_surv * (1 - events[i] / at_risk[i])
  }
  survival_prob[i] <- current_surv
}

# Create Kaplan-Meier table
km_table <- data.frame(
  Time = unique_times,
  Events = events,
  Censored = censored,
  AtRisk = at_risk,
  SurvivalProb = survival_prob
)
# Print the first few rows
print(km_table, row.names = FALSE)
```

```
##  Time Events Censored AtRisk SurvivalProb
##   1.8      1        0     25       0.9600
##   2.2      1        0     24       0.9200
##   2.5      1        0     23       0.8800
##   2.6      1        0     22       0.8400
##   3.0      0        1     21       0.8400
##   3.5      1        0     20       0.7980
##   3.8      1        0     19       0.7560
##   5.3      1        0     18       0.7140
##   5.4      1        0     17       0.6720
##   5.7      0        1     16       0.6720
##   6.6      1        0     15       0.6272
##   8.2      1        0     14       0.5824
##   8.7      1        0     13       0.5376
##   9.2      2        0     12       0.4480
##   9.8      1        0     10       0.4032
##  10.0      0        1      9       0.4032
##  10.2      1        0      8       0.3528
##  10.7      1        0      7       0.3024
##  11.0      1        0      6       0.2520
##  11.1      1        0      5       0.2016
##  11.7      1        0      4       0.1512
##  12.1      0        1      3       0.1512
##  12.2      0        1      2       0.1512
##  12.3      0        1      1       0.1512
```

3. The corresponding Kaplan-Meier survival curve would be,

```
plot(km_table$Time, km_table$SurvivalProb, type = "s", col = "blue", lwd = 2,
     xlab = "Time (Months)", ylab = "Survival Probability",
     main = "Kaplan-Meier Survival Curve", ylim = c(0, 1))

grid()
```



```
# Also

# plot(survfit(Surv(time, status) ~  1, type ="kaplan-meier", data =dat),
#       conf.int =False)
```

3 The Cox Proportional Hazards Model and Its Characteristics

3.1 Exercise 4

The following table shows the results of a Cox regression analysis to study the effect of hormonal therapy on recurrence free survival time in 686 breast cancer patients.

Variable	Coef.	Haz. Ratio	Std. Err.	z	$p > \|z\|$	lower 95% CI	upper 95% CI
Hormonal therapy	-0.346	?	0.129	-2.683	0.007	0.549	?
Age	-0.009	0.991	0.009	?	?	0.973	1.009
Menopausal status = post	?	1.295	0.183	?	0.159	0.904	?
Tumor grade = L	0.551	1.736	0.190	2.904	?	?	?
Tumor grade = Q	-0.201	0.818	0.122	-1.649	0.099	0.644	1.039
Number of positive nodes	0.049	1.050	0.007	6.551	?	?	1.065
Progesterone receptor	-0.002	?	0.001	-3.866	0.0001	0.997	0.999
Estrogen receptor	0.0002	1.0002	?	0.438	0.661	0.999	1.001

1. Complete the missing information.
2. Based on these results, how would you interpret the association between hormonal therapy and recurrence-free survival in breast cancer patients?
3. Why are the parameters of the Cox PH model fitted based on a partial likelihood?
4. Why would it be inappropriate to use a linear or logistic regression for survival data?

Answer:

1. Completing the Missing Information
   (a) Hormonal Therapy: Hazard Ratio (HR): $e^{-0.346} \approx 0.707$ and Upper 95% CI: $e^{-0.346+1.96\times0.129} \approx 0.911$
   (b) Age : z-statistic: $\frac{-0.009}{0.009} = -1.0$ and $p > |z|$: 0.317
   (c) "Menopausal Status = Post": Coefficient: $\ln(1.295) \approx 0.258$ and z-statistic: $\frac{0.258}{0.183} \approx 1.41$ and Upper 95% CI: $e^{0.258+1.96\times0.183} \approx 1.852$
   (d) "Tumor Grade = L": $p > |z|$: 0.004 and Lower 95% CI: $e^{0.551-1.96\times0.190} \approx 1.196$ and Upper 95% CI: $e^{0.551+1.96\times0.190} \approx 2.517$
   (e) Number of Positive Nodes: $p > |z|$: 5.7e-11 and Lower 95% CI: $e^{0.049-1.96\times0.007} \approx 1.036$
   (f) Progesterone Receptor: Hazard Ratio: $e^{-0.002} \approx 0.998$
   (g) Estrogen Receptor: Standard Error: $\frac{0.0002}{0.438} \approx 0.0005$

2. The **hazard ratio (HR) for hormonal therapy is 0.707**, indicating a decreases by a factor of 0.707 e.g. women who do not receive the hormonal therapy have a hazard that $\frac{1}{0.71} = 41\%$ higher than those women who received the therapy. Hormonal therapy shows a significant association with recurrence free survival time.

3. Cox PH model parameters fitted based on a "partial" likelihood

   - **Key feature of the Cox model**: It does not assume a distribution for the outcome variable (time-to-event).

   - Unlike parametric models, the Cox PH model **cannot use a full likelihood** based on the outcome distribution.

   - The likelihood is constructed using the **observed order of events** rather than their joint distribution.

   - **Censored observations** contribute indirectly by remaining in the risk set for subsequent failures but are not explicitly included in likelihood terms. They apear in the risk set of the terms for each failure.

4. Linear or logistic regression for survival data

   - Both linear and logistic regression ignore censoring, a critical aspect of survival data.

   - **Example of logistic regression limitation** (George et al., 2014):
     Two treatment groups had identical final event proportions. One group experienced events **immediately after randomization**, while the other had events **just before follow-up ended**. Despite identical proportions, the treatments have **different clinical implications** due to event timing. Logistic regression fails to account for **time-to-event**, leading to incorrect conclusions.

3.2 Exercise 5

The following tables present results on the association between survival and prehospital administration of aspirin and heparin in patients without out-of-hospital cardiac arrest. The variable Treatment (1 = aspirin/heparin administration, 0 = no aspirin/heparin administration) indicates the exposure of interest. Gender (1 = male, 0 = female) and Diagnosis may act as confounding or effect-modifying variables. The Diagnosis variable takes the value 1 for patients diagnosed with myocardial infarction and 0 for other diagnoses, including pulmonary embolism, primary arrhythmia, and aortic rupture.

Model 1

Variable	Coef.	Haz. Ratio	Std. Err.	p
Treatment	-0.730	0.482	0.141	2.5e-07

Log-Likelihood = -1956.960

Model 2

Variable	Coef.	Haz. Ratio	Std. Err.	p
Treatment	-0.727	0.484	0.142	2.9e-07
Gender	-0.126	0.882	0.126	0.32

Log-Likelihood = -1956.476

Model 3

Variable	Coef.	Haz. Ratio	Std. Err.	p
Treatment	-0.755	0.470	0.302	0.01
Gender	-0.132	0.877	0.138	0.34
Treatment*Gender	0.037	1.037	0.342	0.91

Log-Likelihood = -1956.470

Model 4

Variable	Coef.	Haz. Ratio	Std. Err.	p
Treatment	-0.429	0.651	0.146	0.0034
Diagnosis	-0.979	0.376	0.119	<2e-16

Log-Likelihood = -1920.849

Model 5

Variable	Coef.	Haz. Ratio	Std. Err.	p
Treatment	-0.062	0.940	0.189	0.743
Diagnosis	-0.826	0.438	0.129	1.6e-10
Treatment*Diagnosis	-0.755	0.470	0.287	0.009

Log-Likelihood = -1917.405

1. Give the mathematical expression of Model 1 to Model 5.

2. Based on the parameter estimates in Model 2, calculate the hazard ratio between a woman who received aspirin and a man who did not receive aspirin.

3. Based on the parameter estimates in Model 5, calculate the hazard ratio between a patient who was diagnosed with a myocardial infarction and who received aspirin and another patient who was diagnosed with a pulmonary embolism and did not receive aspirin.

4. How would you characterize the influence of gender and diagnosis in the association between aspirin/heparin treatment and survival (confounding effect, effect modification, both, none)?

Answer:

1. Each model follows the Cox proportional hazards model:

$$h(t|X) = h_0(t)\exp(\beta_1 X_1 + \beta_2 X_2 + \dots)$$

where:

- $h(t|X)$ is the hazard function at time t given covariates X.
- $h_0(t)$ is the baseline hazard function.
- X_i are covariates (e.g., treatment, gender, diagnosis).
- β_i are the corresponding coefficients.

Model	**Specification**
Model 1	$h(t \mid \text{Treatment}) = h_0(t)\exp(\beta_1 \cdot \text{Treatment})$
Model 2	$h(t \mid \text{Treatment, Gender}) = h_0(t)\exp(\beta_1 \cdot \text{Treatment} + \beta_2 \cdot \text{Gender})$
Model 3	$h(t \mid \text{Treatment, Gender}) = h_0(t)\exp(\beta_1 \cdot \text{Treatment} + \beta_2 \cdot \text{Gender} + \beta_3 \cdot (\text{Treatment} \times \text{Gender}))$
Model 4	$h(t \mid \text{Treatment, Diagnosis}) = h_0(t)\exp(\beta_1 \cdot \text{Treatment} + \beta_2 \cdot \text{Diagnosis})$
Model 5	$h(t \mid \text{Treatment, Diagnosis}) = h_0(t)\exp(\beta_1 \cdot \text{Treatment} + \beta_2 \cdot \text{Diagnosis} + \beta_3 \cdot (\text{Treatment} \times \text{Diagnosis}))$

2. For Model 2, Calculating HR between a **woman who received aspirin** and a **man who did not receive aspirin**

$$HR = \frac{e^{1\times -0.727 + 0\times -0.126}}{e^{0\times -0.727 + 1\times -0.126}} = \frac{e^{-0.727}}{e^{-0.126}} = 0.548.$$

Therefore, The woman who received aspirin has **0.548 times the hazard** of the man who did not receive aspirin.

3. In Model 5, calculating HR between **a myocardial infarction patient who received aspirin** and **a pulmonary embolism patient who did not receive aspirin**:

$$HR = \frac{e^{1\times -0.062 + 1\times -0.826 + 1\times -0.755}}{e^{0\times -0.062 + 0\times -0.826 + 0\times -0.755}} = \frac{e^{-1.643}}{e^{0}} = 0.193.$$

Which means that the myocardial infarction patient who received aspirin has **0.193 times the hazard** compared to the pulmonary embolism patient who did not receive aspirin.

---

4. Confounding vs. Effect Modification

Gender: Neither confounding variable nor effect modification.

In Model 2, when gender is added to the model, the hazard ratio for treatment remains nearly the same compared to Model 1 (0.484 vs. 0.482). The effect of gender itself is not statistically significant (p = 0.32), indicating that gender does not confound the relationship between treatment and survival. In Model 3, the interaction term (Treatment × Gender) is also not significant (p = 0.91), suggesting that gender does not modify the effect of treatment on survival. Since neither the inclusion of gender alone nor its interaction with treatment significantly changes the treatment effect, gender is neither a confounder nor an effect modifier in this analysis.

Variable	**Coef**	**Haz. Ratio**	**Std. Err.**	**P value**
Model 1	**Treatment**	−0.730	0.482	0.141
Model 2	**Treat.**	−0.727	0.484	0.142
	Gender	0.126	0.882	0.126
Model 3	**Treat.**	0.755	0.470	0.302
	Gender	−0.132	0.877	0.138
	Treat.*Gend.	0.037	1.037	0.342

Diagnosis: Confounding and effect modification.

In Model 4, the hazard ratio for treatment changes substantially from 0.482 in Model 1 to 0.651 in Model 4 after adjusting for diagnosis. The p-value for diagnosis is highly significant (p < 2e-16), indicating that diagnosis strongly affects survival. The change in the treatment effect after adjusting for diagnosis suggests that diagnosis is a confounder in the relationship between treatment and survival. In Model 5, the interaction term (Treatment × Diagnosis) is statistically significant (p = 0.009), implying that the effect of treatment on survival depends on the diagnosis. This indicates effect modification. Since diagnosis influences both the treatment effect and survival while also interacting with treatment, it acts as both a confounder and an effect modifier.

Variable	**Coef**	**Haz. Ratio**	**Std. Err.**	**P value**
Model 1	**Treatment**	−0.730	0.482	0.141
Model 4	**Treat.**	−0.429	0.651	0.146
	Diagn.	−0.979	0.376	0.119
Model 5	**Treat.**	−0.062	0.940	0.189
	Diagn.	−0.826	0.438	0.129
	Treat.*Diagn.	−0.755	0.470	0.287

4 Evaluating the Proportional Hazards Assumption

4.1 Exercise 6

In this exercise, we consider the AIDS Clinical Trials Group (ACTG320) randomized double-blind, placebo-controlled study, which aimed to examine the effectiveness of a new three-drug treatment regimen (including the drug indinavir) when compared to the standard two-drug regimen in HIV-infected patients. The outcome of interest is time to AIDS diagnosis or death in days. The following table shows the results of a Cox regression analysis including the variables treatment (1 = treatment, 0 = control), ivdrug (intravenous drug use history; 0 = never, 1 = ever), karnof (Karnofsky Performance Scale with three categories 100, 90, 80/70), age (age at enrollment), and cd4 (baseline CD4 count).

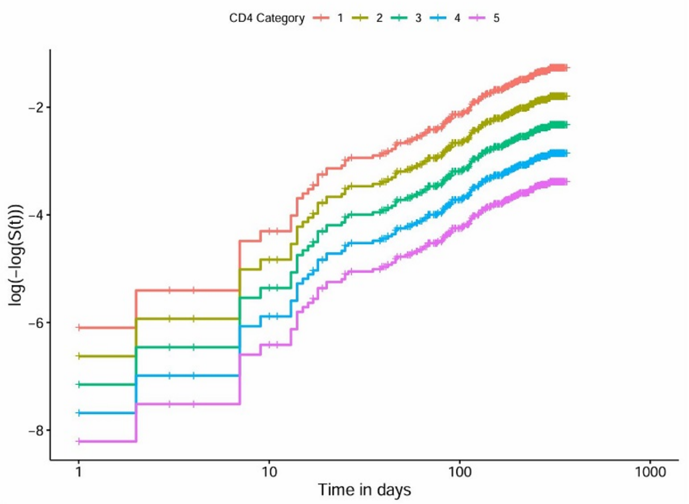
Variable	Coef	Haz. Ratio	Std. Error	z	$p > \|z\|$	$p(PH)$
treatment	-0.67	0.51	0.22	-3.12	0.00	0.41
ivdrug	-0.53	0.59	0.32	-1.66	0.10	0.99
karnof 70-80	1.20	3.34	0.29	4.09	0.00	0.28
karnof 90	0.43	1.53	0.29	1.46	0.14	0.28
age	0.02	1.02	0.01	2.04	0.04	0.26
cd4	-0.01	0.99	0.01	-5.80	0.00	0.12

1. Formulate the proportional hazards assumption underlying the presented Cox model.

2. The table above also presents the result of a goodness-of-fit test for the PH assumption based on Schoenfeld residuals. How do you interpret the result? What are the advantages and drawbacks of using such a GOF test for checking the PH assumption?

3. Download the ACTG data set from the Moodle page (actg320.txt) and plot the log-log survival curves for the categorized CD4 variable cd4 group (the five categories correspond to the observed cd4 quintiles), e.g., using the ggsurvplot function from the survminer package. What do you conclude about whether or not cd4 satisfies the PH assumption? What are the advantages and drawbacks of this graphical approach compared to the GOF test?

4. Consider a Cox regression as shown in the table above but with cd4 group instead of cd4. The following plot shows the log-log survival curves for each cd4 group category adjusted for treatment, ivdrug, karnof, and age, which were obtained by substituting the values for each category of cd4 group and overall mean values for the remaining variables into the formula for the estimated survival curve that results from the Cox regression. Why are these curves parallel and why is this plot not suited to evaluate the PH assumption? Is there a way to still make use of these curves to check the PH assumption?



5. The table below presents the results of a stratified Cox model with cd4 group as a stratification variable. Why is the coefficient for cd4 group missing? How do the hazard ratios change between the strata?

Variable	Coef	Haz. Ratio	Std. Error	z	$p > \|z\|$
treatment	-0.68	0.51	0.22	-3.16	0.00
ivdrugd	-0.54	0.58	0.32	-1.68	0.09
karnof 70-80	1.27	3.55	0.29	4.30	0.00
karnof 90	0.45	1.56	0.29	1.53	0.13
age	0.02	1.02	0.01	2.02	0.04

6. Fit a stratified Cox model with cd4 group as a stratification variable that includes an interaction between cd4 group and treatment. Compare this model to the no-interaction stratified Cox model (i.e., the model that corresponds to the table shown in exercise 6.5) using a likelihood ratio test.

Answer:

1. The proportional hazards assumption in the Cox model states that the hazard ratio (HR) for each covariate remains constant over time. Mathematically:

$$h(t|X) = h_0(t)\exp(\beta_1\text{treatment} + \beta_2\text{ivdrug} + \beta_3\text{karnof} + \beta_4\text{age} + \beta_5\text{cd4})$$

Where:

- $h_0(t)$ = baseline hazard (varies with time)
- Covariates modify hazard multiplicatively through exp()
- PH assumption = constant covariate effects over time

2. Key points:

- All p-values > 0.05 → No significant PH violations
- Advantages: Formal testing, easy implementation
- Drawbacks: Misses subtle effects, low power in small samples

3. Log-Log Survival Curves for CD4 Group

- Parallel curves → PH assumption satisfied
- Crossing curves → PH violation

4. Adjusted Log-Log Curves Parallelism

Why parallel? - Model-enforced PH assumption - Shows theoretical fit, not actual data

Better alternatives: - Schoenfeld residual plots - Time-dependent coefficients

5. Stratified Cox Model & Missing CD4 Coefficient

Key implications: - Stratification removes CD4 coefficient - Estimates separate baseline hazards - Useful when CD4 violates PH assumption

---

5 The Stratified Cox Procedure

5.1 Exercise 7

The health effects of many exposures can be studied in populations of workers who are exposed to a variety of chemical, biological, or physical (e.g., noise, heat, radiation) agents. The following table shows data concerning the time until the first diagnosis of lung cancer and the occupational exposure to coal dust for five workers in a coal mine. The follow-up for lung cancer started at their first employment and silica exposure is measured in mg per m^3.

Id	Year of Birth	Year of Employment	Silica 2005	Silica 2006	Silica 2007	Year of Diagnosis	Year of Last News
1	1978	1997	0.73	0.38	0.27	NA	2011
2	1959	1985	0.28	0.46	0.00	2007	2014
3	1953	2002	0.35	0.00	1.17	2014	2017
4	1985	2005	0.00	0.00	0.00	NA	2008
5	1954	1982	0.29	0.53	0.47	2011	2019

1. Which time scale would you recommend for this analysis? Which challenges arise when using time-on-study as the time scale? Which challenges arise when using attained age?

2. Using attained age as the time scale, fill in the following table to transform the data to the so-called Counting Process format in order to account for the time-varying nature of exposure. Use the information on yearly exposure to silica to create a time-varying variable `exposure_cumulative` which gives information on cumulative exposure to silica (i.e., silica exposure received until time t).

Id	t_{ij0}	t_{ij1}	exposure_cumulative	d_{ij}
1	19	27	0.00	0
1	27	28	0.73	
1	28			
...				

Answer:

1.
   - Attained age should be preferred as time scale whenever attained age is the stronger determinant of the outcome than time-on-study.
   - Lung cancer mortality is strongly influenced by age and individual 1 who was first employed at age 19 has a much lower risk to die of lung cancer during the follow-up period than individual 3 who was first employed at age 49.
   - At the end of follow-up individual 1 is only 33 years old and the probability to die of lung cancer before age 40 is almost 0.
   - Challenges when ussing attained age: have to account for the resulting left-truncation in the data (for instance, by using a Counting Process format).
   - Challenges when using time-on-study as time scale: have to model the effect of attained age, which is often difficult as this effect will typically not be linear, thereby requiring more flexible modeling strategies.

2. In the Counting Process format multiple lines are used for the same individual and an individual's total at-risk-follow-up time is subdivided into smaller time intervals:

Id	t_{ij0}	t_{ij1}	exposure$_{\text{cumulative}}$	d_{ij}
1	19	27	0.00	0
1	27	28	0.73	0
1	28	29	1.11	0
1	29	33	1.38	0
2	26	46	0.00	0
2	46	47	0.28	0
2	47	48	0.74	1
3	49	52	0.00	0
3	52	54	0.35	0
3	53	61	1.52	1