Chapter 11

Survival Data and the Kaplan-Meier Curve

We have already investigated supervised learning models and hypothesis tests in cases where the outcome of interest is a category or number. But what if the outcome is a *time duration*? For example, what if we're comparing the effects of two treatments and our outcome is the time between treatment administration and disease progression?

Data where the outcome is a time duration are very common in clinical data science and are called **time-to-event** data or **survival data**. The field of **survival analysis** develops methods to analyze and interpret such data. We will examine one such method today and many more in subsequent chapters.

11.1 Example: Ovarian Cancer Survival Dataset

Today we'll examine some data from a study of ovarian cancer[1]. The dataset contains information on 26 women. The variables are:

---

[1] The dataset comes from the `survival` package in R and is labeled `ovarian`. The original study is Edmonson JH *et al*, "Different chemotherapeutic sensitivities and host factors affecting prognosis in advanced ovarian carcinoma versus minimal residual disease", *Cancer Treatment Reports*, 63(2): 241-247; 1979.

- `futime`: The number of days from enrollment in the study until death or censoring, whichever came first

- `fustat`: An indicator of death (1) or censoring (0)

- `age`: The patient's age in years at the time of treatment administration

- `resid.ds`: Residual disease present at the time of treatment administration (1 = no, 2 = yes)

- `rx`: Treatment group (1 = cyclophosphamide, 2 = cyclophosphamide + adriamycin)

- `ecog.ps`: A measure of performance score or functional status at the time of treatment administration, using the Eastern Cooperative Oncology Group's (ECOG) scale. It ranges from 0 (fully functional) to 4 (completely disabled). Level 4 subjects are usually considered too ill to enter a randomized trial such as this. The patients in this dataset are all at Levels 1 and 2.

Here is a histogram of the follow-up times (`futime`) in days, colored according to whether the patient died or was censored (`fustat`):
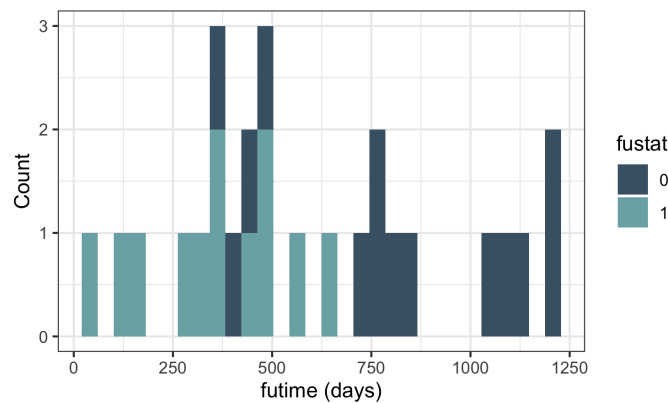
And here is the same graph colored by treatment group (`rx`):
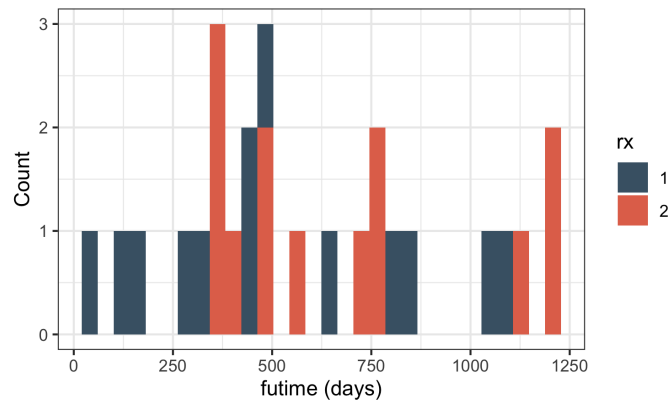
Now, imagine that we want to study the effect of the treatment group (`rx`) on the outcome of death or no death (1 = death, 0 = no death). We could think of this as a classification problem with only a single feature: treatment group. Unfortunately, this method of analyzing time-dependent data is fraught with problems:

1. How do you choose the time horizon at which to evaluate mortality?

2. How do you handle people who dropped out of the study before that time?

11.2   Definitions

Censoring occurs when the event of interest in a time-to-event analysis is not observed. It is a form of missing data problem (see Chapter **??**) and can be caused by a variety of factors, including inconsistencies in follow-up, the study's ending before all subjects have experienced the event, or a lack of knowledge about when, exactly, the event occurred. The type of censoring represented in the `ovarian` dataset is called **right-censoring**. We will focus on right-censoring today and investigate other types later.

> **Right censoring:** A situation that arises when the event of interest has not occurred by the end of the follow-up period. This may be because (a) the study itself ends, (b) a patient is lost to follow-up

during the study period, or (c) a patient experiences a different event that makes further follow-up impossible[2].

Survival data are generally described using two probabilities, called the survival and hazard.

Survival: Also called the **survival function** or **survival probability** and abbreviated $S(t)$, this is the probability that an individual survives to time t (i.e., does not experience the event by time t).

Hazard: Usually denoted by $h(t)$ or $\lambda(t)$, this is the probability that an individual who has not yet experienced the event at time t experiences it at that exact time. In other words, it is the instantaneous event rate for an individual who has already survived to time t.

We will focus on the survival function now and learn more about the hazard later.

11.3   The Kaplan-Meier Estimator

The **Kaplan-Meier estimator** is a nonparametric estimate of the survival function, usually represented graphically by a **Kaplan-Meier curve**[3]. The Kaplan-Meier estimator looks like this:

$$\hat{S}(t) = \prod_{j|t_j \leq t} \frac{n_j - d_j}{n_j}$$

where d_j is the number of subjects who fail at time t_j and n_j is the number of subjects at risk just prior to t_j. Here is a Kaplan-Meier curve for the `ovarian` dataset. The little "+" signs correspond to censoring events.

---

[2]For more information, please see Clark TG *et al*, "Survival Analysis Part I: Basic Concepts and First Analyses", *British Journal of Cancer*, 89, 232–238; 2003.

[3]It can be shown mathematically that the Kaplan-Meier estimator is the maximum likelihood estimator (see Chapter 5) of the survival function in the case of censoring.
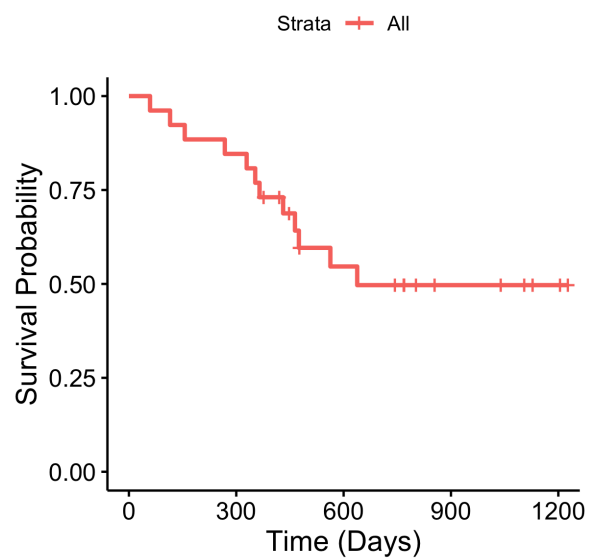
And here are Kaplan-Meier curves for the two treatment groups separately:
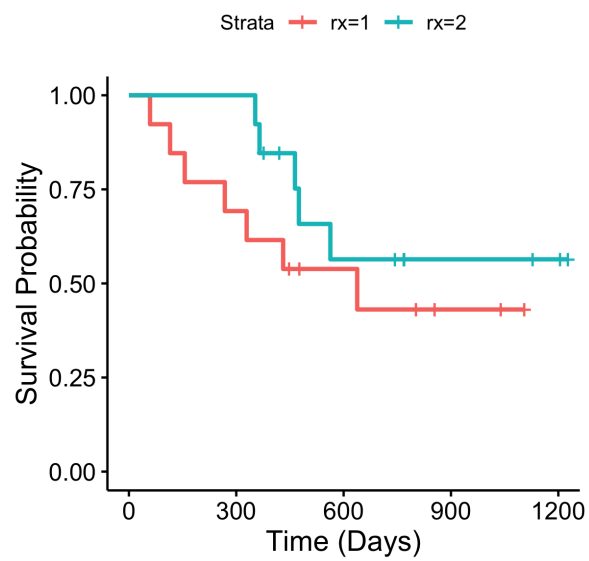
Question 11.1

Here are the raw data from treatment group 1 of the `ovarian` dataset. Using these data, fill in the remaining cells of the table below.

	rx	futime	fustat
1	1	59	1
2	1	115	1
3	1	156	1
4	1	268	1
5	1	329	1
6	1	431	1
7	1	448	0
8	1	477	0
9	1	638	1
10	1	803	0
11	1	855	0
12	1	1040	0
13	1	1106	0

j	t_j	n_j	d_j	$\hat{S}(t_j)$	Calculation
0	0	13	0	1.000	$\frac{13-0}{13}$
1	59	13	1	0.923	$\hat{S}(t_0)\left(\frac{13-1}{13}\right)$
2	115	12	1	0.846	$\hat{S}(t_1)\left(\frac{12-1}{12}\right)$
3	156				
4	268				
5	329	9	1	0.615	$\hat{S}(t_4)\left(\frac{9-1}{9}\right)$
6	431	8	1	0.538	$\hat{S}(t_5)\left(\frac{8-1}{8}\right)$
7	448	7	0	0.538	$\hat{S}(t_6)\left(\frac{7-0}{7}\right)$
8	477	6	0	0.538	$\hat{S}(t_7)\left(\frac{6-0}{6}\right)$
9	638	5	1	0.431	$\hat{S}(t_8)\left(\frac{5-1}{5}\right)$
10	803	4	0		
11	855	3	0		
12	1040	2	0		
13	1106	1	0		

11.4 Assumptions of the Kaplan-Meier Estimator

The Kaplan-Meier estimator makes three important assumptions:

1. The probability of censoring is unrelated to the outcome of interest.

2. The survival probabilities are the same for participants recruited at different times during the study (e.g., circumstances that could alter the survival, such as treatments, do not change over calendar time).

3. The events occurred at exactly the times specified.

11.5 Comparing Kaplan-Meier Curves

Of course, now the question arises: How do we formally compare two Kaplan-Meier curves? There is a nonparametric hypothesis test for comparing Kaplan-Meier curves called the log-rank test; we will see it in Chapter **??**. There is also an entire family of linear models, called Cox proportional hazards models, that use the Kaplan-Meier curve as their backbone and model the effects of different covariates on this curve. We will see them in Chapter **??**.

	rx	futime	fustat
1	2	353	1
2	2	365	1
3	2	377	0
4	2	421	0
5	2	464	1
6	2	475	1
7	2	563	1
8	2	744	0
9	2	769	0
10	2	770	0
11	2	1129	0
12	2	1206	0
13	2	1227	0