11.3 the Kaplan-Meier Survival Curve

The Survival Curve

- The survival function (or curve) is defined as

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

- This decreasing function quantifies the probability of surviving past time t.
- Ex: suppose a company is interested in modeling customer churn. Let T be the time that a customer cancels a subscription.
- Then $S(t)$ represents the probability a customer cancels later than time t.
- The larger the value of $S(t)$, the less likely the customer will cancel before time t

Estimating the Survival Curve

- Consider the `BrainCancer` dataset, which contains the survival times for patients with primary brain tumors undergoing treatment with stereotactic radiation methods.
- The predictors are `gtv` (gross tumor volume, in cubic centimeters); `sex` (male or female); `diagnosis` (meningioma, LG glioma, HG glioma, or other); `loc` (the tumor location: either infratentorial or supratentorial); `ki` (Karnofsky index); and `stereo` (stereotactic method).
- 53 of the 88 patients were still alive at the end of the study.

BrainCancer Data

```
attach(BrainCancer)
table(sex)
```

```
## sex
## Female   Male
##     45     43
```

```
table(diagnosis)
```

```
## diagnosis
## Meningioma  LG glioma  HG glioma      Other
##         42          9         22         14
```

```
table(status) # status = 1 indicates an uncensored observation
```

```
## status
##  0  1
## 53 35
```

Estimating the Survival Curve

- Suppose we'd like to estimate $S(20) = P(T > 20)$, the probability that a patient survives for at least 20 months.
- It is tempting to compute the proportion of patients who are known to have survived past 20 months.
  - This turns out to be 48/88, or approximately 55%.
- However, 17 of the 40 patients who did not survive to 20 months were actually censored, and this analysis implicitly assumes they died before 20 months. So it is probably an underestimate.

Estimating the Survival Curve

- Let $d_1 < d_2 < \cdots < d_K$ represent the K unique death times among he non-censored patients, and let q_k denote the number of patients who died at time d_k.
- Let r_k denote the number of patients alive and in the study just before d_k; these are the *at risk* patients.
- By the law of total probability,

$$P(T > d_k) = P(T > d_k | T > d_{k-1})P(T > d_{k-1}) + P(T > d_k) = P(T > d_k | T \leq d_{k-1})P$$

- Since $d_{k-1} < d_k$, $P(T > d_k | T \leq d_{k-1}) = 0$ and we can write

$$S(d_k) = P(T > d_k | T > d_{k-1})P(T > d_{k-1})$$

Estimating the Survival Curve

Reworking this a bit, we find

$$S(d_k) = P(T > d_k | T > d_{k-1}) P(T > d_{k-1} | T > d_{k-2}) \times \cdots \times P(T > d_2 | T > d_1) P(T > d_1)$$

and we just need to plug in estimates of each of the terms.

We use the estimator

$$\hat{P}(T > d_j | T > d_{j-1}+ = (r_j - q_j)/r_j$$

the fraction of the risk set at time d_j who survived past time d_j.

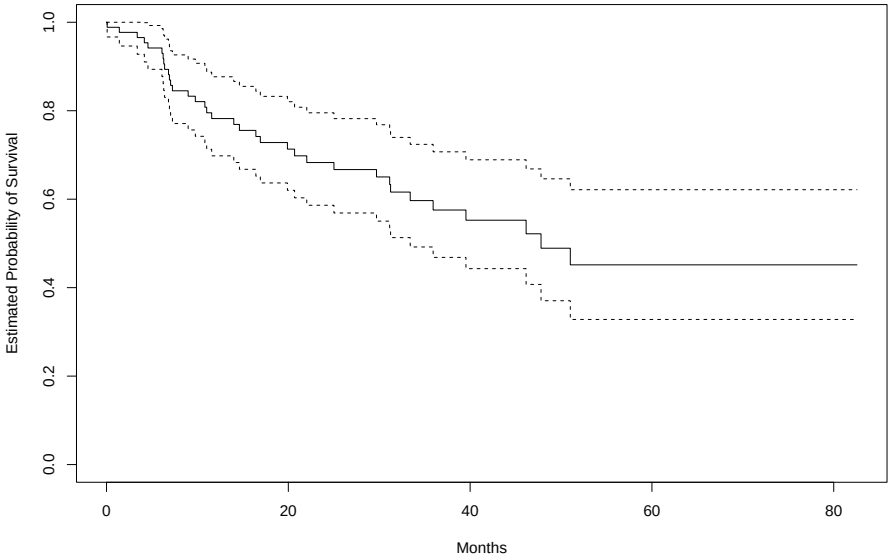
The Kaplan-Meier Estimator

The Kaplan-Meier Estimator of the survival curve is

$$\hat{S}(d_k) = \prod_{j=1}^{k} \left(\frac{r_j - q_j}{r_j} \right)$$

Example: Brain Cancer Data

```
library(survival)
fit.surv <- survfit(Surv(time, status) ~ 1)
plot(fit.surv, xlab = "Months",
    ylab = "Estimated Probability of Survival")
```

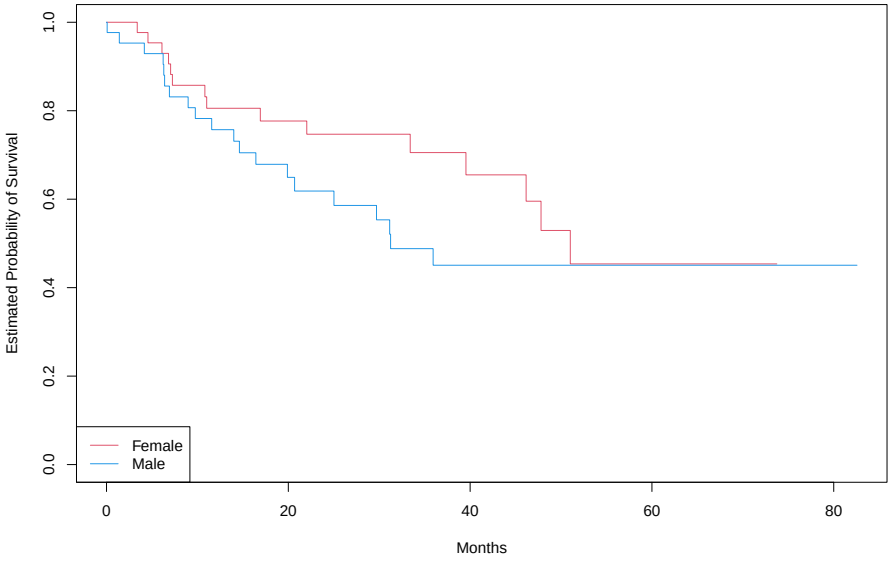

Estimated Probability of Survival
1.0
0.8
0.6
0.4
0.2
0.0
0
20
40
60
80
Months

Example: Brain Cancer Data

- Each point in the solid step-like curve shows the estimated probability of surviving past the time indicated on the horizontal axis.
- The estimated probability of survival past 20 months is 71%, which is quite a bit higher than the naive estimate of 55% presented earlier.

Example: stratify by sex

```
fit.sex <- survfit(Surv(time, status) ~ sex)
plot(fit.sex, xlab = "Months",
    ylab = "Estimated Probability of Survival", col = c(2,4))
legend("bottomleft", levels(sex), col = c(2,4), lty = 1)
```


Estimated Probability of Survival
1.0
0.8
0.6
0.4
0.2
0.0
Female
Male
0
20
40
60
80
Months