Lecture 15
Introduction to Survival Analysis

BIOST 515

February 26, 2004

Background

In logistic regression, we were interested in studying how risk factors were associated with presence or absence of disease. Sometimes, though, we are interested in how a risk factor or treatment affects time to disease or some other event. Or we may have study dropout, and therefore subjects who we are not sure if they had disease or not. In these cases, logistic regression is not appropriate.

Survival analysis is used to analyze data in which the time until the event is of interest. The response is often referred to as a **failure time, survival time,** or **event time**.

Examples

- Time until tumor recurrence
- Time until cardiovascular death after some treatment intervention
- Time until AIDS for HIV patients
- Time until a machine part fails

The survival time response

- Usually continuous
- May be incompletely determined for some subjects
  - i.e.- For some subjects we may know that their survival time was at least equal to some time t. Whereas, for other subjects, we will know their exact time of event.
- Incompletely observed responses are **censored**
- Is always ≥ 0.

Analysis issues

- If there is no censoring, standard regression procedures could be used.

- However, these may be inadequate because
  - Time to event is restricted to be positive and has a skewed distribution.
  - The probability of surviving past a certain point in time may be of more interest than the expected time of event.
  - The hazard function, used for regression in survival analysis, can lend more insight into the failure mechanism than linear regression.

Censoring

Censoring is present when we have some information about a subject's event time, but we don't know the exact event time. For the analysis methods we will discuss to be valid, censoring mechanism must be independent of the survival mechanism.

There are generally three reasons why censoring might occur:

- A subject does not experience the event before the study ends
- A person is lost to follow-up during the study period
- A person withdraws from the study

These are all examples of **right-censoring**.

Types of right-censoring

- **Fixed type I censoring** occurs when a study is designed to end after C years of follow-up. In this case, everyone who does not have an event observed during the course of the study is censored at C years.

- In **random type I censoring**, the study is designed to end after C years, but censored subjects do not all have the same censoring time. This is the main type of right-censoring we will be concerned with.

- In **type II** censoring, a study ends when there is a pre-specified number of events.

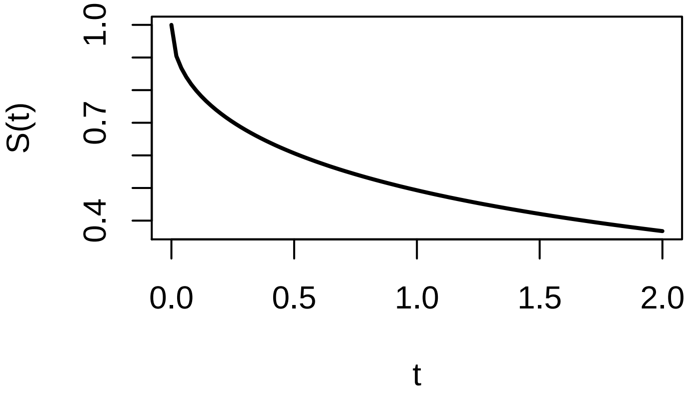
Regardless of the type of censoring, we must assume that it is **non-informative** about the event; that is, the censoring is caused by something other than the impending failure.

Terminology and notation

- T denotes the response variable, $T \geq 0$.

- The survival function is

$$S(t) = Pr(T > t) = 1 - F(t).$$

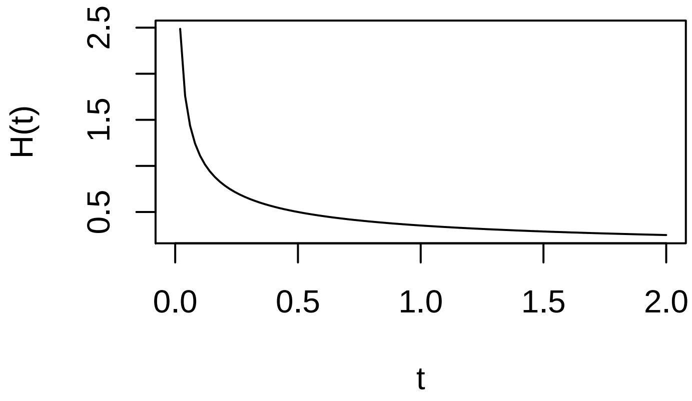


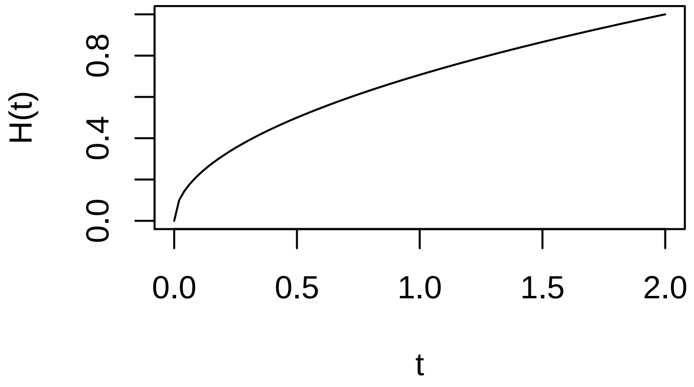
- The survival function gives the probability that a subject will survive past time t.
- As t ranges from 0 to ∞, the survival function has the following properties
  - It is non-increasing
  - At time $t = 0$, $S(t) = 1$. In other words, the probability of surviving past time 0 is 1.
  - At time $t = \infty$, $S(t) = S(\infty) = 0$. As time goes to infinity, the survival curve goes to 0.
- In theory, the survival function is smooth. In practice, we observe events on a discrete time scale (days, weeks, etc.).

- The hazard function, $h(t)$, is the instantaneous rate at which events occur, given no previous events.

$$h(t) = \lim_{\Delta t \to 0} \frac{Pr(t < T \leq t + \Delta t | T > t)}{\Delta t} = \frac{f(t)}{S(t)}.$$

- The cumulative hazard describes the accumulated risk up to time t, $H(t) = \int_0^t h(u)du$.





If we know any one of the functions $S(t)$, $H(t)$, or $h(t)$, we can derive the other two functions.

$$h(t) = -\frac{\partial \log(S(t))}{\partial t}$$

$$H(t) = -\log(S(t))$$
$$S(t) = \exp(-H(t))$$

Survival data

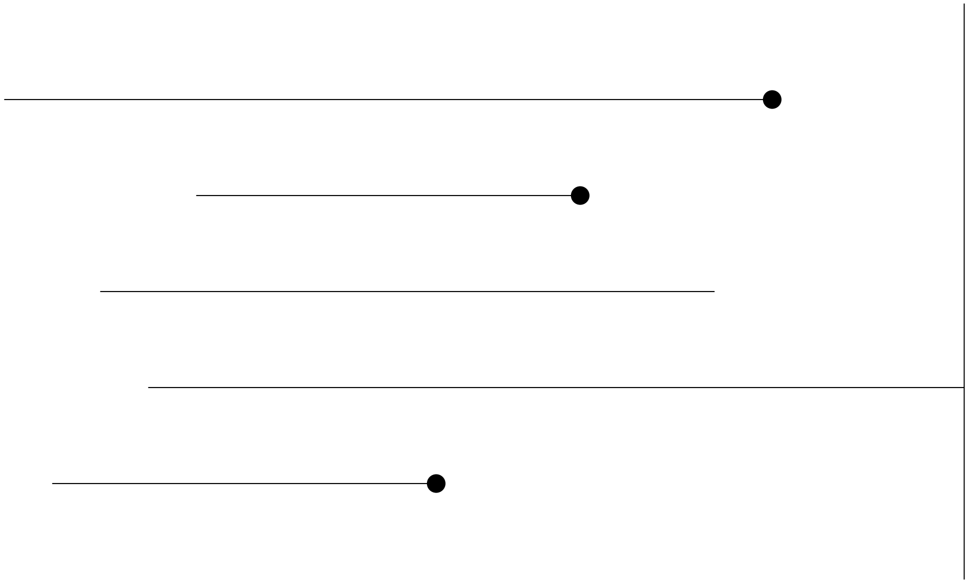
How do we record and represent survival data with censoring?

- T_i denotes the response for the ith subject.
- Let C_i denote the censoring time for the ith subject
- Let δ_i denote the event indicator

$$\delta_i = \begin{cases} 1 \text{ if the event was observed } (T_i \leq C_i) \\ 0 \text{ if the response was censored } (T_i > C_i). \end{cases}$$

- The observed response is $Y_i = \min(T_i, C_i)$.

Example



T_i	C_i	Y_i	δ_i
80	100	80	1
40	80	40	1
74+	74	74	0
85+	85	85	0
40	95	40	1

Termination of study

Estimating $S(t)$ and $H(t)$

If we are assuming that every subject follows the same survival function (no covariates or other individual differences), we can easily estimate $S(t)$.

- We can use nonparametric estimators like the Kaplan-Meier estimator
- We can estimate the survival distribution by making parametric assumptions
  - exponential
  - Weibull
  - Gamma
  - log-normal

Non-parametric estimation of S

- When no event times are censored, a non-parametric estimator of $S(T)$ is $1 - F_n(t)$, where $F_n(t)$ is the empirical cumulative distribution function.

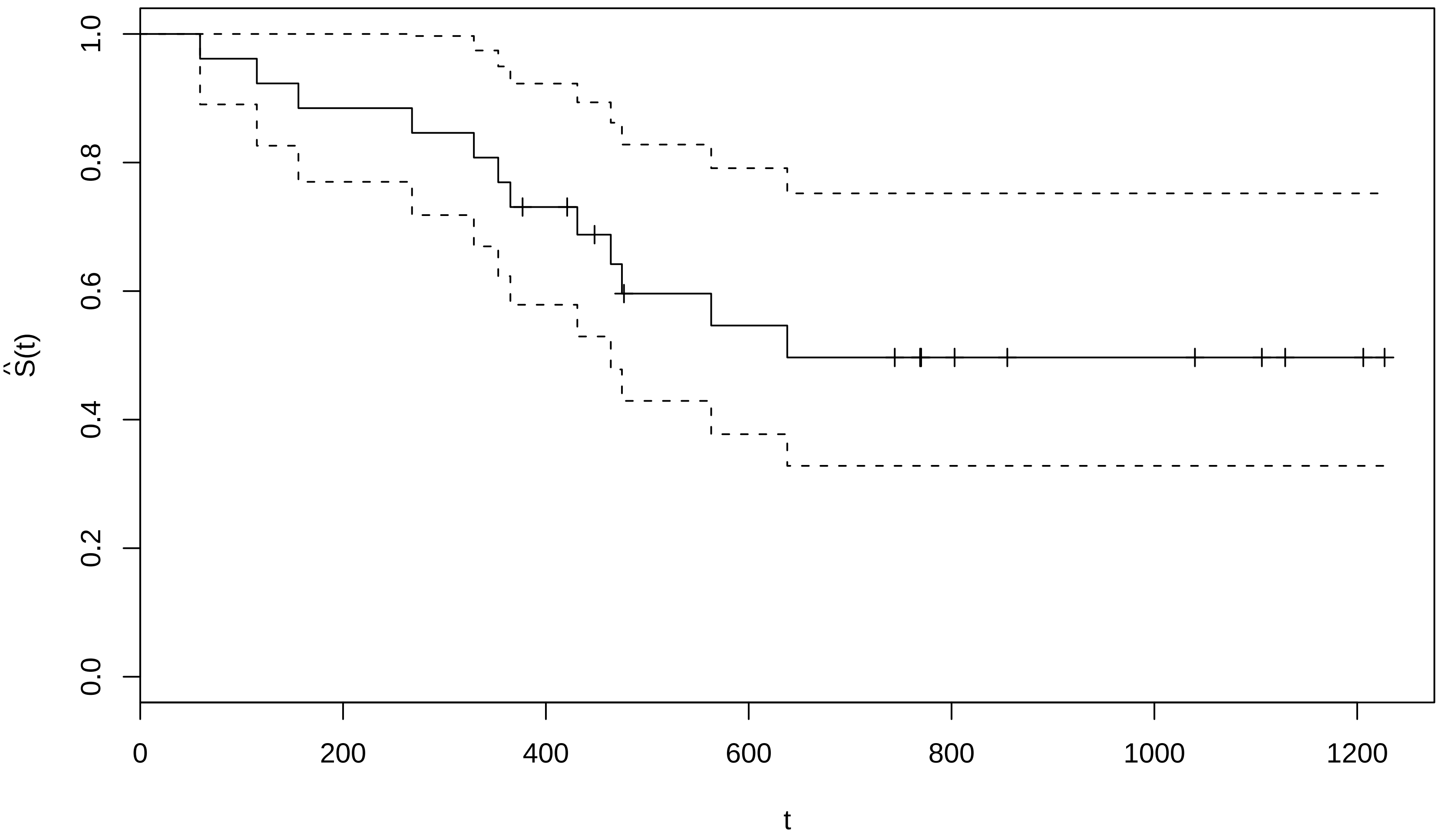
- When some observations are censored, we can estimate $S(t)$ using the Kaplan-Meier product-limit estimator.

t	No. subjects at risk	Deaths	Censored	Cumulative survival
59	26	1	0	$25/26 = 0.962$
115	25	1	0	$24/25 \times 0.962 = 0.923$
156	24	1	0	$23/24 \times 0.923 = 0.885$
268	23	1	0	$22/23 \times 0.885 = 0.846$
329	22	1	0	$21/23 \times 0.846 = 0.808$
353	21	1	0	$20/21 \times 0.808 = 0.769$
365	20	0	1	$20/20 \times 0.769 = 0.769$
377	19	0	1	$19/19 \times 0.769 = 0.769$
421	18	0	1	$18/18 \times 0.769 = 0.769$
431	17	1	0	$16/17 \times 0.769 = 0.688$
$\vdots$				$\vdots$
$\vdots$				$\vdots$

How can we get this in R?

```
> library(survival)
> data(ovarian)
> S1=Surv(ovarian$futime,ovarian$fustat)
> S1
 [1]   59   115   156   421+  431   448+  464   475   477+  563   638   744+
[13]  769+  770+  803+  855+ 1040+ 1106+ 1129+ 1206+ 1227+  268   329   353
[25]  365   377+
> fit1=survfit(S1)
> summary(fit1)
Call: survfit(formula = S1)

 time n.risk n.event survival std.err lower 95% CI upper 95% CI
   59     26       1    0.962  0.0377        0.890        1.000
  115     25       1    0.923  0.0523        0.826        1.000
  156     24       1    0.885  0.0627        0.770        1.000
  268     23       1    0.846  0.0708        0.718        0.997
  329     22       1    0.808  0.0773        0.670        0.974
  353     21       1    0.769  0.0826        0.623        0.949
  365     20       1    0.731  0.0870        0.579        0.923
  431     17       1    0.688  0.0919        0.529        0.894
  464     15       1    0.642  0.0965        0.478        0.862
  475     14       1    0.596  0.0999        0.429        0.828
  563     12       1    0.546  0.1032        0.377        0.791
  638     11       1    0.497  0.1051        0.328        0.752
```

```
>plot(fit1,xlab="t",ylab=expression(hat(S)*"(t)"))
```



Parametric survival functions

The Kaplan-Meier estimator is a very useful tool for estimating survival functions. Sometimes, we may want to make more assumptions that allow us to model the data in more detail. By specifying a parametric form for $S(t)$, we can

- easily compute selected quantiles of the distribution
- estimate the expected failure time
- derive a concise equation and smooth function for estimating $S(t)$, $H(t)$ and $h(t)$
- estimate $S(t)$ more precisely than KM **assuming** the parametric form is correct!

Appropriate distributions

Some popular distributions for estimating survival curves are

- Weibull
- exponential
- log-normal ($\log(T)$ has a normal distribution)
- log-logistic

Estimation for parametric $S(t)$

We will use maximum likelihood estimation to estimate the unknown parameters of the parametric distributions.

- If Y_i is uncensored, the ith subject contributes $f(Y_i)$ to the likelihood
- If Y_i is censored, the ith subject contributes $Pr(y > Y_i)$ to the likelihood.

The joint likelihood for all n subjects is

$$L = \prod_{i:\delta_i=1}^{n} f(Y_i) \prod_{i:\delta_i=0}^{n} S(Y_i).$$

The log-likelihood can be written as

$$\log L = \sum_{i:\delta_i=1}^{n} \log(h(Y_i)) - \sum_{i=1}^{n} H(Y_i).$$

Example

Let's look at the ovarian data set in the $survival$ library in R. Suppose we assume the time-to-event follows an distribution, where

$$h(t) = \lambda$$

and

$$S(t) = \exp(-\lambda t).$$

```
> s2=survreg(Surv(futime, fustat)~1  , ovarian, dist='exponential')
> summary(s2)

Call:
survreg(formula = Surv(futime, fustat) ~ 1, data = ovarian, dist = "exponential"
            Value Std. Error    z         p
(Intercept)  7.17      0.289 24.8 3.72e-136

Scale fixed at 1
```

```
Exponential distribution
Loglik(model)= -98   Loglik(intercept only)= -98
Number of Newton-Raphson Iterations: 4
n= 26
```

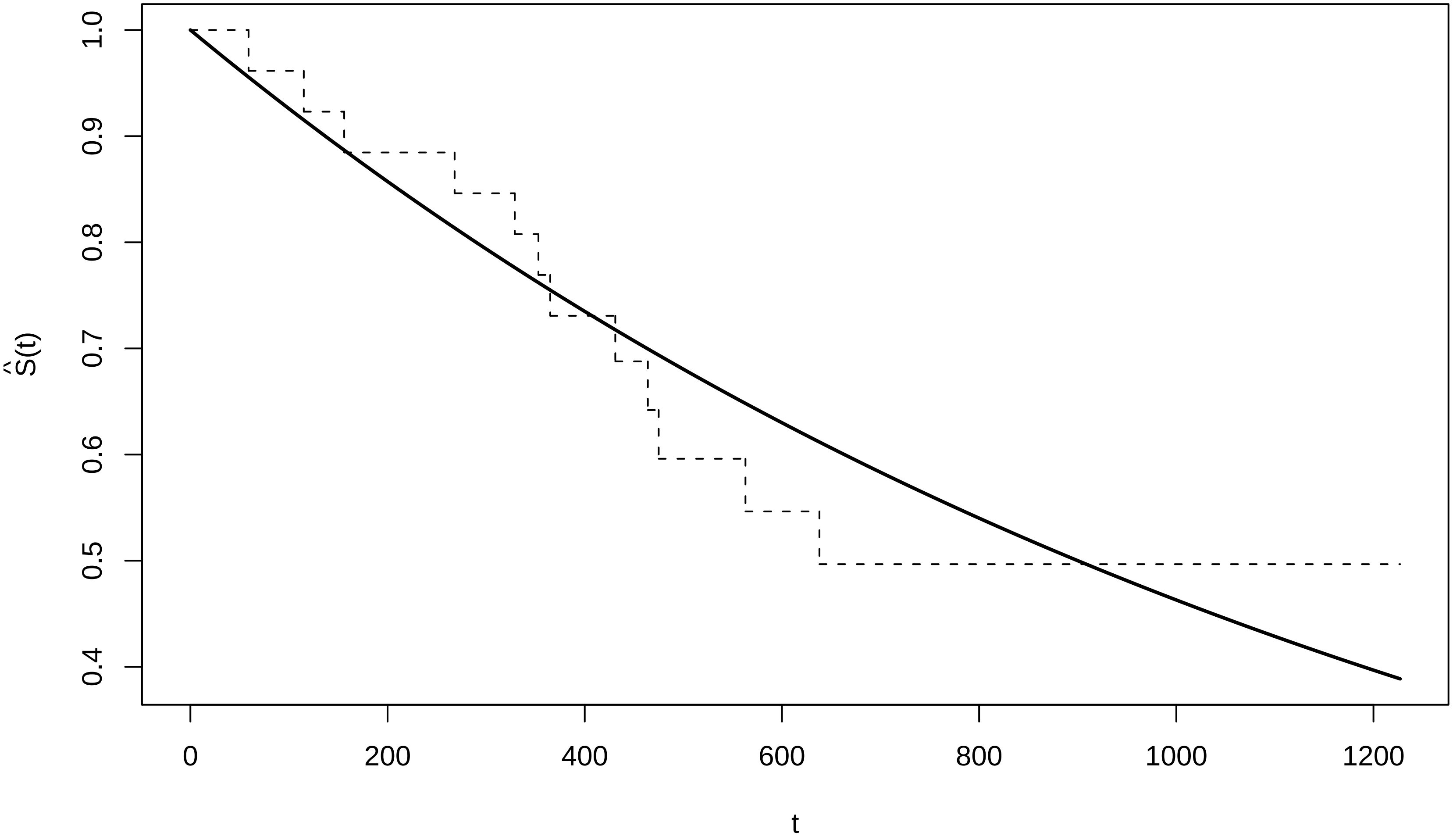
In the R output,

$$\begin{aligned}\lambda &= \exp(-(\text{Intercept}))\\ &= \exp(-7.17)\end{aligned}$$

Therefore,

$$S(t) = \exp(-\exp(-7.17)t).$$

```
plot(T,1-pexp(T,exp(-7.169)),xlab="t",ylab=expression(hat(S)*"(t)"))
```



Example

Let's look at the ovarian data set in the $survival$ library in R. Suppose we assume the time-to-event follows a Weibull distribution, where

$$h(t) = \alpha\gamma t^{\gamma-1}$$

and

$$S(t) = \exp(-\alpha t^{\gamma}).$$

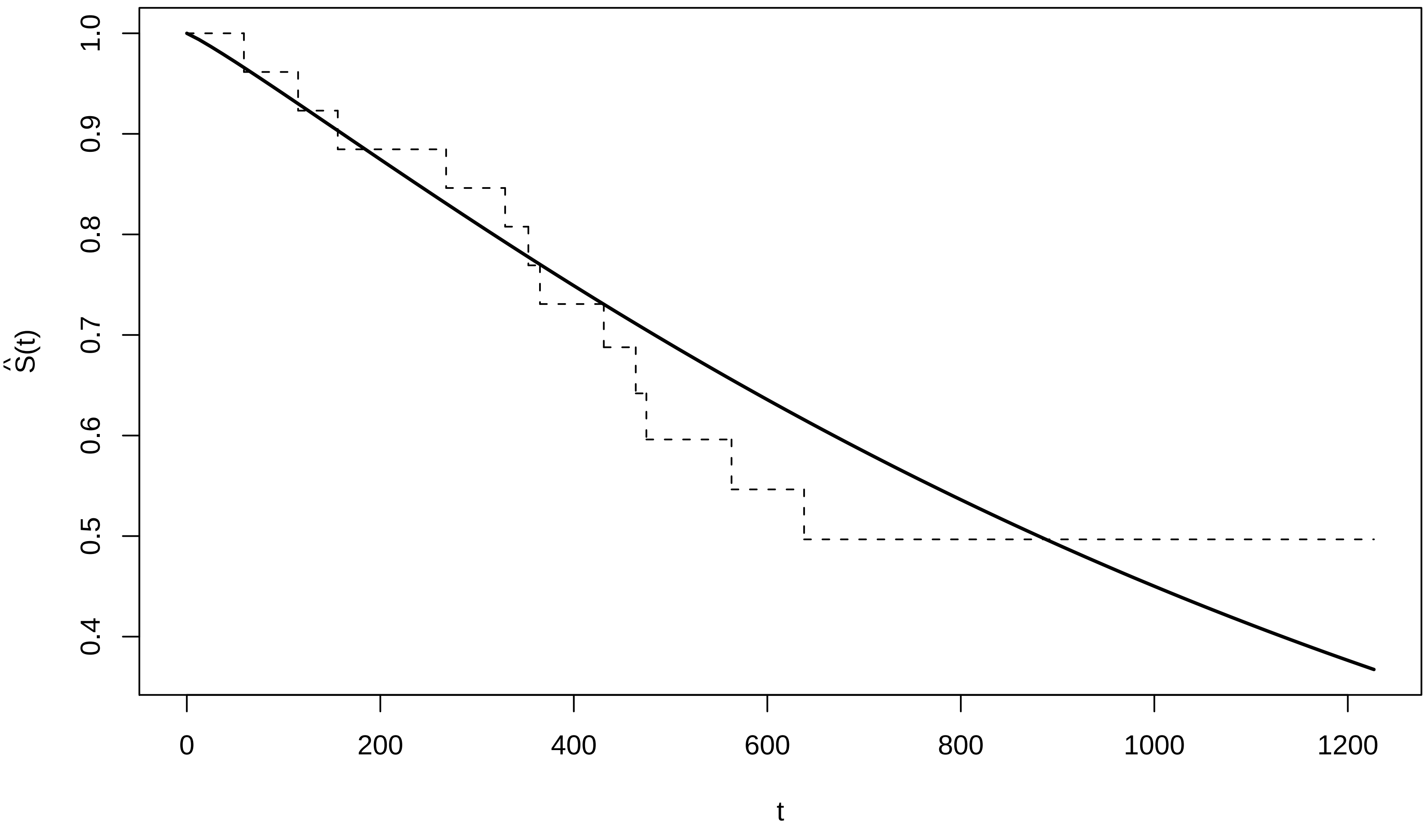
```
> s1=survreg(Surv(futime, fustat)~1  , ovarian, dist='weibull',scale=0)
> summary(s1)

Call:
survreg(formula = Surv(futime, fustat) ~ 1, data = ovarian, dist = "weibull",
    scale = 0)
             Value Std. Error      z         p
(Intercept)  7.111      0.293 24.292 2.36e-130
Log(scale)  -0.103      0.254 -0.405  6.86e-01

Scale= 0.902
```

```
Weibull distribution
Loglik(model)= -98   Loglik(intercept only)= -98
Number of Newton-Raphson Iterations: 5
n= 26
```

To match the notation above, $\gamma = 1/\text{Scale}$ and $\alpha = \exp(-(\text{Intercept})\gamma)$. This gives us the following survival function,

$$S(t) = \exp(-\exp[-7.111/.902])t^{1/.902}).$$

```
plot(T,1-pweibull(T,1/0.902,exp(7.111)))
```


$\hat{S}(t)$
t