Introduction to Survival Analysis

David M. Rocke

April 27, 2021

Time to Event Data

- Survival Analysis is a term for analyzing time-to-event data.
- This is used in clinical trials, where the event is often death or recurrence of disease.
- It is used in engineering reliability analysis, where the event is failure of a device or system.
- It is used in insurance, particularly life insurance, where the event is death.

Time to Event Data

- The distribution of ‘failure’ times is asymmetric and can be long-tailed.
- The base distribution is not normal, but exponential.
- There are usually *censored* observations, which are ones in which the failure time is not observed.

Time to Event Data

- Often, these are *right-censored*, meaning that we know that the event occurred after some known time t, but we don't know the actual event time, as when a patient is still alive at the end of the study.
- Observations can also be *left-censored*, meaning we know the event has already happened at time t, or *interval-censored*, meaning that we only know that the event happened between times t_1 and t_2.
- Analysis is difficult if censoring is associated with treatment.

Right Censoring

- Patients are in a clinical trial for cancer, some on a new treatment and some on standard of care.
- Some patients in each group have died by the end of the study. We know the survival time (measured for example from time of diagnosis—each person on their own clock).
- Patients still alive at the end of the study are right censored.
- Patients who are lost to follow-up or withdraw from the study may be right-censored.

Left and Interval Censoring

- An individual tests positive for HIV.
- If the event is infection with HIV, then we only know that it has occurred before the testing time t, so this is left censored.
- If an individual has a negative HIV test at time t_1 and a positive HIV test at time t_2, then the infection event is interval censored.

Basic Quantities and Models

The probability density function $f(x)$ is defined as with any continuous distribution. For any short interval of time, it can be thought of as the chance that the event will occur in that short interval. The cumulative distribution function is

$$F(x) = \Pr(X \leq x) = \int_0^x f(x)dx$$

For survival data, a more relevant quantity is the *survival function*

$$S(x) = 1 - F(x) = \Pr(X > x) = \int_x^\infty f(x)dx$$

Basic Quantities and Models

$$S(x) = 1 - F(x) = \Pr(X > x) = \int_x^\infty f(x)dx$$

The survival function $S(x)$ is the probability that the event time is later than x. If the event in a clinical trial is death, then this is the fraction of the original population at time 0 that is still alive at time x; that is, the fraction surviving to time x.

The Hazard Function

Another important function is the *hazard function*, which is the probability that the event will occur in the next very short interval, given that it has not occurred yet.

$$h(x) = \lim_{\Delta x \to 0} \frac{\Pr[x \leq X < x + \Delta x | X \geq x]}{\Delta x}$$

The expression in the numerator is the probability of survival until at least time $x + \Delta x$ conditional on surviving until time x. This might be the chance of someone who has just turned 30 still being alive one day later.

The Hazard Function

$$h(x) = \lim_{\Delta x \to 0} \frac{\Pr[x \le X < x + \Delta x | X \ge x]}{\Delta x}$$

This might be the chance of someone who has just turned 30 still being alive one day later. You can see that this is different than the probability at birth of surviving until age 30 plus one day.

The Hazard Function

$$\begin{aligned}
h(x) &= \lim_{\Delta x \to 0} \frac{\Pr[x \le X < x + \Delta x | X \ge x]}{\Delta x} \\
&= S^{-1}(x) \lim_{\Delta x \to 0} \frac{\Pr[x \le X < x + \Delta x]}{\Delta x} \\
&= f(x)/S(x)
\end{aligned}$$

The limit takes the difference quotient into a derivative (by definition of the derivative) and the result is because the density $f(x)$ is the derivative of the CDF $F(x)$.

The Hazard Function

Also,

$$\begin{aligned} h(x) &= \lim_{\Delta x \to 0} \frac{\Pr[x \le X < x + \Delta x | X \ge x]}{\Delta x} \\ &= f(x)/S(x) \\ f(x) &= -\frac{dS(x)}{dx} \\ h(x) &= -\frac{d\ln(S(x))}{dx} \end{aligned}$$

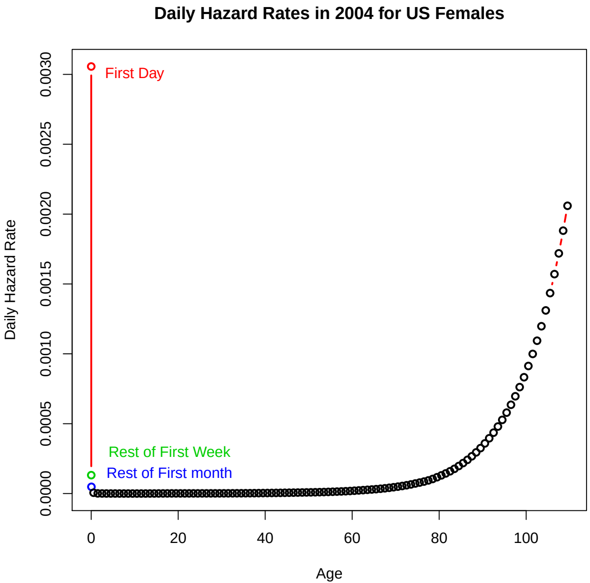
Cumulative Hazard

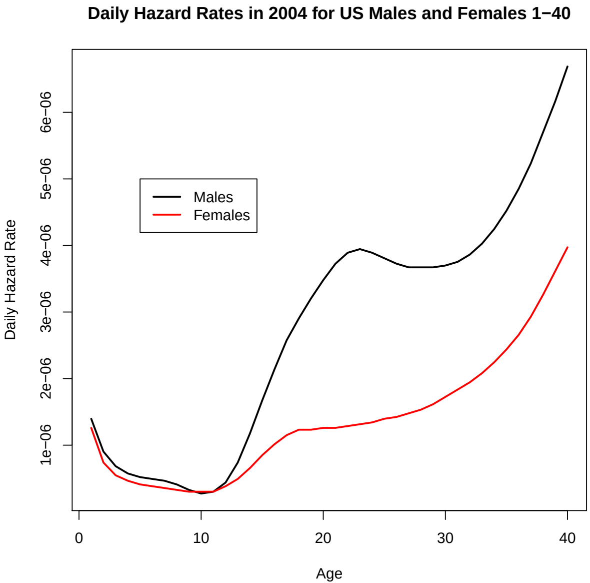
$$h(x) = -\frac{d\ln(S(x))}{dx}$$

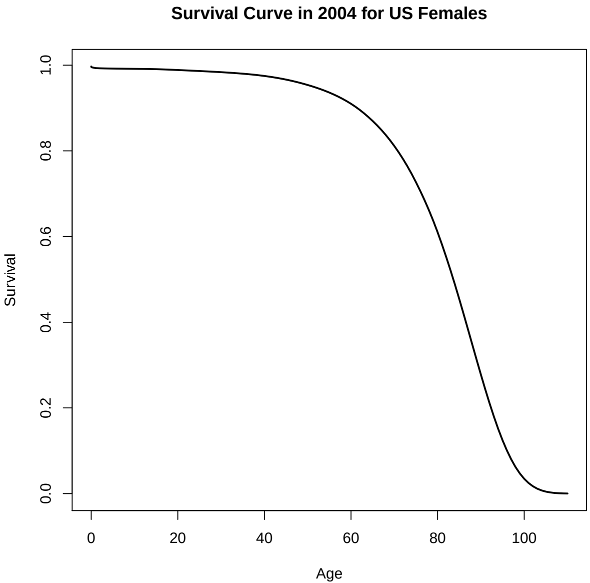
The cumulative hazard function is

$$H(x) = \int_0^x h(x)dx = -\ln(S(x))$$

This function is easier to estimate than the hazard function, and we can then approximate the hazard function by the approximate derivative of the cumulative hazard.


Daily Hazard Rates in 2004 for US Females
Daily Hazard Rate
0.0000
0.0005
0.0010
0.0015
0.0020
0.0025
0.0030
First Day
Rest of First Week
Rest of First month
0
20
40
60
80
100
Age


Daily Hazard Rates in 2004 for US Males and Females 1−40
Males
Females
Daily Hazard Rate
1e−06
2e−06
3e−06
4e−06
5e−06
6e−06
0
10
20
30
40
Age


Survival Curve in 2004 for US Females
Survival
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
100
Age

Exponential Distribution

- The exponential distribution is the base distribution for survival analysis.
- The distribution has a constant hazard λ
- The mean survival time is λ^{-1}

$$
\begin{aligned}
f(x) &= \lambda e^{-\lambda x} \\
\ln(f(x)) &= \ln \lambda - \lambda x \\
F(x) &= 1 - e^{-\lambda x} \\
S(X) &= e^{-\lambda x} \\
\ln(S(x)) &= -\lambda x \\
h(x) &= -\frac{d}{dx} \ln(S(x)) \\
&= -\frac{d}{dx}(-\lambda x) \\
&= \lambda
\end{aligned}
$$

Estimation of λ

- Suppose we have m exponential survival times of $t_1, t_2, \ldots, t_m$ and k right-censored values at $u_1, u_2, \ldots, u_k$.
- A survival time of $t_i = 10$ means that subject i died at time 10. A right-censored time $u_i = 10$ means that at time 10, subject i was still alive and that we have no further follow-up.
- For the moment we will assume that the survival distribution is exponential and that all the subjects have the same parameter λ.

Estimation of λ

We have m exponential survival times of $t_1, t_2, \ldots, t_m$ and k right-censored values at $u_1, u_2, \ldots, u_k$. The log-likelihood of an observed survival time t_i is

$$\ln\left(\lambda e^{-\lambda t_i}\right) = \ln\lambda - \lambda t_i$$

and the likelihood of a censored value is the probability of that outcome (survival greater than u_j) so the log-likelihood is

$$\log(\lambda e^{u_j}) = -\lambda u_j.$$

Let $T = \sum t_i$ and $U = \sum u_j$. Then the log likelihood is

$$\sum_{i=1}^{m}(\ln\lambda - \lambda t_i) + \sum_{j=1}^{k}(-\lambda u_j) = m\ln\lambda - (T+U)\lambda$$

$$m \ln \lambda - (T + U)\lambda$$

is maximized when the derivative wrt λ is 0, that is when

$$\begin{aligned} 0 &= m/\hat{\lambda} - (T + U) \\ \hat{\lambda} &= m/(T + U) \\ 1/\hat{\lambda} &= (T + U)/m \end{aligned}$$

Thus, the estimated mean survival is the total of the times, exact and censored, divided by the number of exact times. It can be show that the variance of $\hat{\lambda}$ is asymptotically λ^2/m, depending only on the number of uncensored observations. This is generally true.

The Score and the Fisher Information

The log likelihood is

$$\ell\ell = m \ln \lambda - (T + U)\lambda$$

and its derivative, called the *score*, is

$$\ell\ell' = m/\lambda - (T + U)$$

Under certain conditions, the negative derivative of the score, called the *Fisher Information*, estimates the reciprocal of the variance of the MLE.

The score is

$$\ell\ell' = m/\lambda - (T + U)$$

(which is 0 evaluated at the MLE) and the observed Fisher information is

$$-\ell\ell'' = m/\hat{\lambda}^2$$

and its reciprocal is

$$\hat{\lambda}^2/m$$

Although the value of $\hat{\lambda}$ depends on the censored data, the variance depends only on the uncensored sample size.

- The (expected value of the) score statistic is zero when evaluated at the MLE.
- the larger the second derivative of the log likelihood is, the steeper the fall-off from the MLE and the more certainly we know the true parameter.
- The multivariate generalization of the Fisher information is most times the method of determining the variance covariance matrix for Wald tests.

Multivariate Generalization

- If there are p parameters, then the score is the gradient vector of length p of partial derivatives of the log likelihood. This determines the estimates by solving p equations in p unknowns setting the score vector to the zero vector.
- The Hessian H is the matrix of second partials and its negative inverse evaluated at the MLE's estimates the variance covariance matrix of the estimated parameters.

Mean Residual Life

The mean lifetime with a survival distribution $f(x)$ is

$$\int_0^\infty xf(x)dx$$

For the exponential distribution we know that the mean is λ^{-1} The mean residual life after survival to time x is

$$mrl(x) \; = \; \int_x^\infty (u-x)f(u)du/S(x)$$

For the exponential, the mean residual life is also λ^{-1}

- The 2017 US standard mortality table estimates the expectation of life of females at birth as 80.96 years.
- At age 50, 95.5% of US females are still alive.
- The mean residual life at age 50 is 33.23 years (age $50 + 33.23 = 83.23$). At age 83, 55.3% are still alive.
- In 1850 an estimate of the expectation of life at birth for females is 39.4 years. At age 1, it is $1 + 49.3 = 50.3$
- But 44.7% of females lived to age 50 and the further expectation of life was 20.4 years, so to age 70.4. About 24% lived to age 70 and 10% to age 80.
- So it was not rare to live beyond age 39.

Other Parametric Survival Distributions

- Any density on $[0, \infty)$ can be a survival distribution, but the most useful ones are all skew right.
- The commonest generalization of the exponential is the Weibull.
- Other common choices are the gamma, log-normal, log-logistic, Gompertz, inverse Gaussian, and Pareto.
- Most of what we do going forward is non-parametric or semi-parametric, but sometimes these parametric distributions provide a useful approach.

Weibull Distribution

$$\begin{aligned} f(x) &= \alpha\lambda x^{\alpha-1}e^{-\lambda x^{\alpha}} \\ h(x) &= \alpha\lambda x^{\alpha-1} \\ S(x) &= e^{-\lambda x^{\alpha}} \\ E(X) &= \Gamma(1+1/\alpha)/\lambda^{1/\alpha} \end{aligned}$$

When $\alpha = 1$ this is the exponential. When $\alpha > 1$ the hazard is increasing and when $\alpha < 1$ the hazard is decreasing. This provides more flexibility than the exponential.

Nonparametric Survival Analysis

- Mostly, we work without a parametric model.
- The first task is to estimate a survival function from data listing survival times, and censoring times for censored data.
- For example one patient may have relapsed at 10 months. Another might have been followed for 32 months without a relapse having occurred (censored).
- The minimum information we need for each patient is a time and a censoring variable which is 1 if the event occurred at the indicated time and 0 if this is a censoring time.

KM drug6mp Data

This is from a clinical trial in 1963 for 6-MP treatment vs. placebo for Acute Leukemia in 42 children. Pairs of children matched by remission status at the time of treatment (1 = partial or 2 = complete) and randomized to 6-MP or placebo. Followed until relapse or end of study. All of the placebo group relapsed, but some of the 6-MP group were censored (which means they were still in remission).

6-MP = 6-Mercaptopurine (Purinethol) is an anti-cancer ("antineoplastic" or "cytotoxic") chemotherapy drug used currently for Acute lymphoblastic leukemia (ALL). It is classified as an antimetabolite.

KM drug6mp Data

Clinical trial in 1963 for 6-MP treatment vs. placebo for Acute Leukemia in 42 children. Pairs of children matched by remission status at the time of treatment (1 = partial or 2 = complete) and randomized to 6-MP or placebo. Followed until relapse or end of study. All of the placebo group relapsed, but some of the 6-MP group were censored.

```
> library(KMsurv)
> data(drug6mp)
> drug6mp
   pair remstat t1 t2 relapse
1     1       1  1 10       1
2     2       2 22  7       1
3     3       2  3 32       0
```

KM drug6mp Data

```
drug6mp data

Description

The drug6mp data frame has 21 rows and 5 columns.

Format

This data frame contains the following columns:

pair         pair number
remstat      Remission status at randomization (1=partial, 2=complete)
t1           Time to relapse for placebo patients, months
t2           Time to relapse for 6-MP patients, months
relapse      Relapse indicator (0=censored, 1=relapse) for 6-MP patients
```

Descriptive Statistics

- The average time in each group is not useful. Some of the 6-MP patients have not relapsed at the time recorded, while all of the placebo patients have relapsed.
- The median time is not really useful either because so many of the 6-MP patients have not relapsed (12/21).
- Both are biased down in the 6-MP group. Remember that lower times are worse since they indicate sooner recurrence.

Descriptive Statistics

- We can compute the average hazard rate, which is the estimate of the exponential parameter: number of relapses divided by the sum of the times.
- For the placebo, that is just the reciprocal of the mean time $= 1/8.667 = 0.115$.
- For the 6-MP group this is $9/359 = 0.025$
- The estimated average hazard in the placebo group is 4.6 times as large (if the hazard is constant over time).

The Kaplan-Meier Product Limit Estimator

- The estimated survival function for the placebo patients is easy to compute. For any time t in months, $S(t)$ is the fraction of patients with times greater than t.
- For the 6-MP patients, we cannot ignore the censored data because we know that the time to relapse is greater than the censoring time.

The Kaplan-Meier Product Limit Estimator

- For any time t in months, we know that 6-MP patients with times greater than t have not relapsed, and those with relapse time less than t have relapsed, but we don't know if patients with censored time less than t have relapsed or not.
- The procedure we usually use is the Kaplan-Meier product-limit estimator of the survival function.

- The Kaplan-Meir estimator is a step function (like the empirical cdf), which changes value only at the event times, not at the censoring times.
- At each event time t, we compute the at-risk group size Y, which is all those observations whose event time or censoring time is at least t.
- If d of the observations have an event time (not a censoring time) of t, then the group of survivors immediately following time t is reduced by the fraction

$$\frac{Y-d}{Y} = 1 - \frac{d}{Y}$$

If the event times are t_i with events per time of d_i $(1 \leq i \leq k)$, then

$$\hat{S}(t) = \prod_{t_i < t} [1 - d_i / Y_i]$$

where Y_i is the set of observations whose time (event or censored) is $\geq t_i$, the group at risk at time t_i.

If there are no censored data, and there are n data points, then just after (say) the third event time

$$\begin{aligned}\hat{S}(t) &= \prod_{t_i<t}[1-d_i/Y_i] \\ &= [\frac{n-d_1}{n}][\frac{n-d_1-d_2}{n-d_1}][\frac{n-d_1-d_2-d_3}{n-d_1-d_2}] \\ &= \frac{n-d_1-d_2-d_3}{n}\end{aligned}$$

the usual empirical cdf estimate.

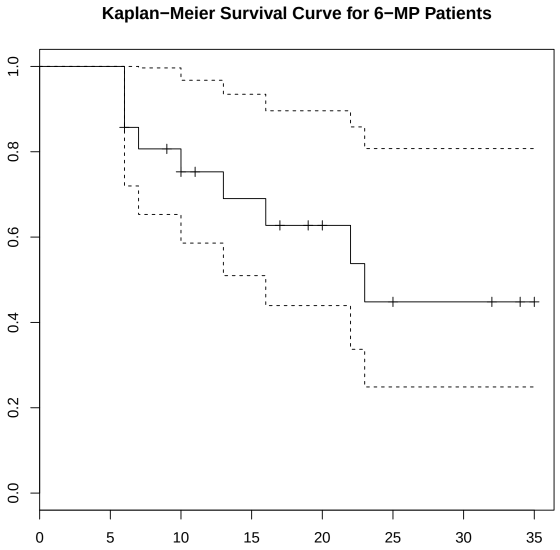
```
require(KMsurv)
data(drug6mp)
plot(survfit(Surv(drug6mp$t2,drug6mp$relapse)~1))
title("Kaplan-Meier Survival Curve for 6-MP Patients")

time12 <- c(drug6mp$t1,drug6mp$t2)
cens12 <- c(rep(1,21),drug6mp$relapse)
treat12 <- rep(1:2,each=21)
pairs12 <- rep(1:21,2)

plot(survfit(Surv(time12,cens12)~treat12),col=1:2)
title("Kaplan-Meier Survival Curve for 6-MP and Placebo Patients")

plot(survfit(Surv(time12,cens12)~treat12),conf.int=T,col=1:2)
title("Kaplan-Meier Survival Curve for 6-MP and Placebo Patients")
```

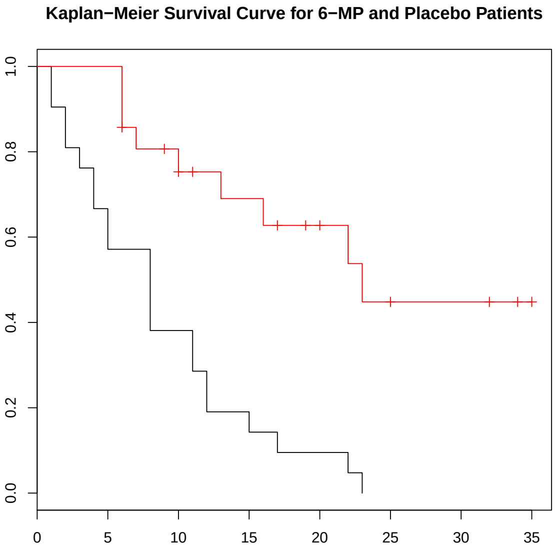
Time	At Risk	Relapses	Censored	KM Factor	KM Curve
6	21	3	1	0.857	0.857
7	17	1	0	0.941	0.807
9	16	0	1	1	0.807
10	15	1	1	0.933	0.753
11	13	0	1	1	0.753
13	12	1	0	0.917	0.690
16	11	1	0	0.909	0.627
17	10	0	1	1	0.627
19	9	0	1	1	0.627
20	8	0	1	1	0.627
22	7	1	0	0.857	0.538
23	6	1	0	0.833	0.448
25	5	0	1	1	0.448
32	4	0	2	1	0.448
34	2	0	1	1	0.448
35	1	0	1	1	0.448

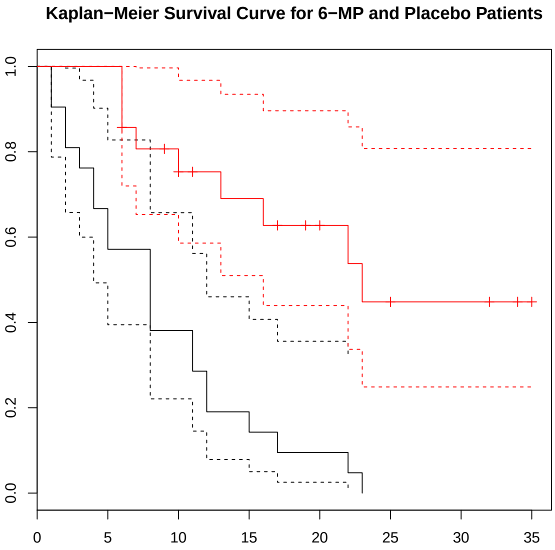
For the 6-MP patients at time 6 months, there are 21 patients at risk. At $t = 6$ there are 3 relapses and 1 censored observations. The Kaplan-Meier factor is $(21 - 3)/21 = 0.857$. The number at risk for the next time $(t = 7)$ is $21 - 3 - 1 = 17$.

At time 7 months, there are 17 patients at risk. At $t = 7$ there is 1 relapse and 0 censored observations. The Kaplan-Meier factor is $(17 - 1)/17 = 0.941$. The Kaplan Meier estimate is $0.857 \times 0.941 = 0.807$. The number at risk for the next time $(t = 9)$ is $17 - 1 = 16$.

```
time12 <- c(drug6mp$t1,drug6mp$t2)
cens12 <- c(rep(1,21),drug6mp$relapse)
treat12 <- rep(1:2,each=21)
pairs12 <- rep(1:21,2)

print(survdiff(Surv(time12,cens12)~treat12))

           N Observed Expected (O-E)^2/E (O-E)^2/V
treat12=1 21       21     10.7      9.77      16.8
treat12=2 21        9     19.3      5.46      16.8

 Chisq= 16.8  on 1 degrees of freedom, p= 4.17e-05

print(survdiff(Surv(time12,cens12)~treat12+strata(pairs12)))

           N Observed Expected (O-E)^2/E (O-E)^2/V
treat12=1 21       21     13.5      4.17      10.7
treat12=2 21        9     16.5      3.41      10.7

 Chisq= 10.7  on 1 degrees of freedom, p= 0.00106
```


Kaplan−Meier Survival Curve for 6−MP Patients
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
25
30
35


Kaplan−Meier Survival Curve for 6−MP and Placebo Patients
1.0
0.8
0.6
0.4
0.2
0.0
0
5
10
15
20
25
30
35


Kaplan−Meier Survival Curve for 6−MP and Placebo Patients
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
25
30
35

Package Survival

```
Surv

Create a survival object, usually used as a response variable in a model formula.

Usage

Surv(time, event)

Arguments

time   for right censored data, this is the follow up time.

event   The status indicator, normally 0=alive, 1=dead.
        Also TRUE/FALSE (TRUE = death) or 1/2 (2=death).
        The event indicator can be omitted,
          in which case all subjects are assumed to have an event.
-----
Surv(drug6mp$t2,drug6mp$relapse)
```

Package Survival

```
survfit

This function creates survival curves from either a formula
(e.g. the Kaplan-Meier), a previously fitted Cox model,
or a previously fitted accelerated failure time model.

Usage

survfit(formula, ...)

Arguments

formula    either a formula or a previously fitted model
-----
plot(survfit(Surv(drug6mp$t2,drug6mp$relapse)~1))
plot(survfit(Surv(time12,cens12)~treat12))
```

Package Survival

```
survdiff

Tests if there is a difference between two or more survival curves.

Usage

survdiff(formula, data, subset, na.action, rho=0)

Arguments

formula      a formula expression as for other survival models,
             of the form Surv(time, status) ~ predictors.
             A strata term may be used to produce a stratified test.

rho          Type of test. Default is the Mantel-Haenszel test.
-------
print(survdiff(Surv(time12,cens12)~treat12))
print(survdiff(Surv(time12,cens12)~treat12+strata(pairs12)))
```