

Survival Analysis

Jos Elkind

March, 2008

1 Survival analysis

2 Application

3 Additional topics

Outline

- 1 Survival analysis
- 2 Application
- 3 Additional topics

Concept

Basic idea is to estimate time until death or failure.

Survivor function

Say, our dependent variable, Y_i , records length of life, thus a random variable between 0 and ∞ .

Survivor function

Say, our dependent variable, Y_i , records length of life, thus a random variable between 0 and ∞ .

The cumulative distribution function of Y is $F(t) = Pr(Y < t)$, i.e. the probability of death younger than t .

Survivor function

Say, our dependent variable, Y_i , records length of life, thus a random variable between 0 and ∞ .

The cumulative distribution function of Y is $F(t) = Pr(Y < t)$, i.e. the probability of death younger than t .

More commonly used is its complement, the probability of death older than t : $S(t) = Pr(Y > t) = 1 - Pr(Y < t) = 1 - F(t)$.

Survivor function

Say, our dependent variable, Y_i , records length of life, thus a random variable between 0 and ∞ .

The cumulative distribution function of Y is $F(t) = Pr(Y < t)$, i.e. the probability of death younger than t .

More commonly used is its complement, the probability of death older than t : $S(t) = Pr(Y > t) = 1 - Pr(Y < t) = 1 - F(t)$.

The latter is known as the **survivor function**.

Survivor function

Say, our dependent variable, Y_i , records length of life, thus a random variable between 0 and ∞ .

The cumulative distribution function of Y is $F(t) = Pr(Y < t)$, i.e. the probability of death younger than t .

More commonly used is its complement, the probability of death older than t : $S(t) = Pr(Y > t) = 1 - Pr(Y < t) = 1 - F(t)$.

The latter is known as the **survivor function**.

Note that $S(0) = 1$ and $S(\infty) = 0$, and $S(t)$ decreases monotonically between 0 and ∞ .

Empirical survivor function

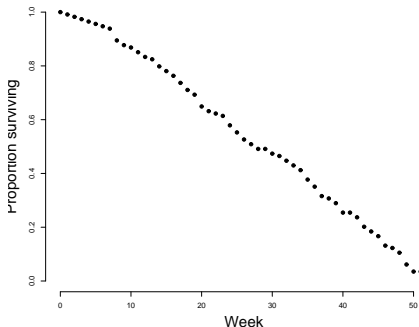
$$S(t) = \frac{\text{number of observations } > t}{n} = \frac{1}{n} \sum_i^n I_{(t, \infty)}(Y_i),$$

whereby $I_{(a,b)}(x)$ is an indicator function which is 1 if x is between a and b , 0 otherwise.

Empirical survivor function

$$S(t) = \frac{\text{number of observations } > t}{n} = \frac{1}{n} \sum_i^n I_{(t, \infty)}(Y_i),$$

whereby $I_{(a,b)}(x)$ is an indicator function which is 1 if x is between a and b , 0 otherwise.



Example: Australian PMs

Open dataset *australian_pm.csv* and plot the empirical survivor function:

```
library(foreign)
apm <- read.csv(file.choose(), header=TRUE)
fix(apm)

plot(survfit(Surv(apm$Months)), conf.int=FALSE)
```

Example: Australian PMs

Open dataset *australian_pm.csv* and plot the empirical survivor function:

```
library(foreign)
apm <- read.csv(file.choose(), header=TRUE)
fix(apm)

plot(survfit(Surv(apm$Months)), conf.int=FALSE)
```

Note that this is in fact an estimated, rather than an empirical plot, but if there is no censoring (see below), the two are equivalent.

Hazard function

We have $S(t) = Pr(Y > t) = 1 - F(t)$. What we are usually interested in is the **hazard function**, the probability of death “now”, $Pr(t < Y < t + \Delta t)$, *given* survival up until now: $Pr(t < Y < t + \Delta t | Y > t)$.

Hazard function

We have $S(t) = Pr(Y > t) = 1 - F(t)$. What we are usually interested in is the **hazard function**, the probability of death “now”, $Pr(t < Y < t + \Delta t)$, given survival up until now: $Pr(t < Y < t + \Delta t | Y > t)$.

As Δt goes to 0, this is given by:

$$h(t) = \frac{f(t)}{S(t)} = \frac{f(t)}{1 - F(t)}$$

Hazard function

We have $S(t) = Pr(Y > t) = 1 - F(t)$. What we are usually interested in is the **hazard function**, the probability of death “now”, $Pr(t < Y < t + \Delta t)$, given survival up until now: $Pr(t < Y < t + \Delta t | Y > t)$.

As Δt goes to 0, this is given by:

$$h(t) = \frac{f(t)}{S(t)} = \frac{f(t)}{1 - F(t)}$$

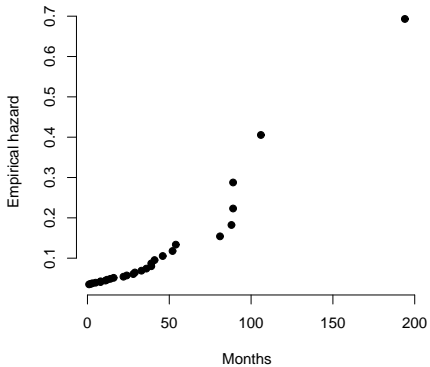
Alternative names: force of decrement, force of mortality, age-specific death (failure) rate, intensity function, hazard function.

Hazard function: derivation

$$\begin{aligned}Pr(t < Y < t + \Delta t | Y > t) &= \frac{Pr(t < Y < t + \Delta t, Y > t)}{Pr(Y > t)} \\ &= \frac{Pr(t < Y < t + \Delta t)}{Pr(Y > t)} \\ \lim_{\Delta t \rightarrow 0} \frac{1}{\Delta t} \times \frac{Pr(t < Y < t + \Delta t)}{Pr(Y > t)} &= \lim_{\Delta t \rightarrow 0} \frac{1}{\Delta t} \times \frac{F(t + \Delta t) - F(t)}{S(t)} \\ &= \frac{F'(t)}{S(t)} \\ &= \frac{f(t)}{S(t)}\end{aligned}$$

Empirical hazard function

$$H(t) = -\log_e S(t)$$



Constant hazard model

The probability of survival is independent of the time of survival thus far.

Constant hazard model

The probability of survival is independent of the time of survival thus far.

$$h(t) = \frac{1}{\beta}$$

Constant hazard model

The probability of survival is independent of the time of survival thus far.

$$h(t) = \frac{1}{\beta}$$

$$S(t) = e^{-\frac{t}{\beta}}$$

Constant hazard model

The probability of survival is independent of the time of survival thus far.

$$h(t) = \frac{1}{\beta}$$

$$S(t) = e^{-\frac{t}{\beta}}$$

The result is an **exponential probability model**.

Weibull hazard model

Another typical model is the Weibull specification:

$$h(t) = \frac{\beta}{\alpha} \left(\frac{t}{\alpha}\right)^{\beta-1}$$

Weibull hazard model

Another typical model is the Weibull specification:

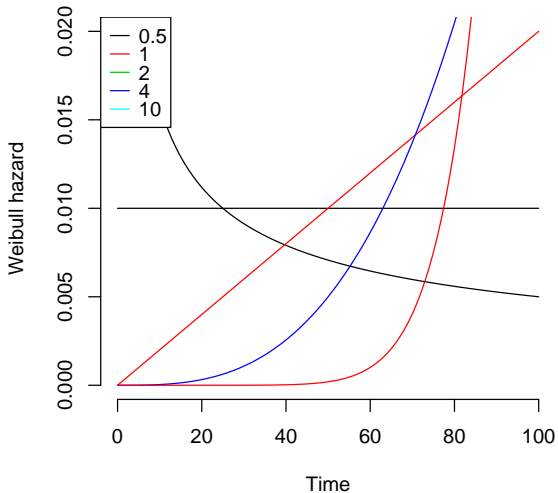
$$h(t) = \frac{\beta}{\alpha} \left(\frac{t}{\alpha}\right)^{\beta-1}$$

If $\beta = 1$, this reduces to:

$$h(t) = \frac{1}{\alpha} \left(\frac{t}{\alpha}\right)^{1-1} = \frac{1}{\alpha},$$

which is the exponential model.

Weibull hazard model



Censoring

Right censoring:

- Cases die for other reasons
- Cases die outside observed timeframe

Censoring

Right censoring:

- Cases die for other reasons
- Cases die outside observed timeframe

Left censoring:

- Birth time is unknown

Censoring

Right censoring:

- Cases die for other reasons
- Cases die outside observed timeframe

Left censoring:

- Birth time is unknown

Note that if censoring is not independent of Y , estimates can be seriously biased.

Censoring in R

A dependent variable for survival analysis is defined in R by the function *Surv*.

Censoring in R

A dependent variable for survival analysis is defined in R by the function *Surv*.

Template 1:

```
Surv(time, event)
```

Where *time* refers to the length of time until death and *event* is 0 when right-censored, 1 when not.

Censoring in R

A dependent variable for survival analysis is defined in R by the function *Surv*.

Template 1:

```
Surv(time, event)
```

Where *time* refers to the length of time until death and *event* is 0 when right-censored, 1 when not.

Template 2:

```
Surv(time, time2, event, type)
```

Where *time* is the interval $[time, time2]$, *type* is “interval” and *event* is 0 for right-censored, 1 for normal event, 2 for left-censored, and 3 for both left- and right-censored.

Proportional hazard model

Generally, we want to estimate survival models with independent, explanatory variables. The typical structure for this is:

$$h_x(t) = h_0(t)g(x) = h_0(t)e^{X\beta}$$

Proportional hazard model

Generally, we want to estimate survival models with independent, explanatory variables. The typical structure for this is:

$$h_x(t) = h_0(t)g(x) = h_0(t)e^{X\beta}$$

$h_0(t)$ is called the **base hazard**.

Proportional hazard model

Generally, we want to estimate survival models with independent, explanatory variables. The typical structure for this is:

$$h_x(t) = h_0(t)g(x) = h_0(t)e^{X\beta}$$

$h_0(t)$ is called the **base hazard**.

Note that the relative hazard of two cases is independent of the base hazard:

$$\frac{h_{x_1}(t)}{h_{x_2}(t)} = \frac{h_0(t)g(x_1)}{h_0(t)g(x_2)} = \frac{g(x_1)}{g(x_2)}$$

Proportional hazard model

The base hazard can now have various different distributions, e.g. exponential.

Proportional hazard model

The base hazard can now have various different distributions, e.g. exponential.

In R:

```
summary(survreg(Surv(y,c) ~ x1 + x2 + x3,  
               data=data, dist="exponential"))
```

```
summary(survreg(Surv(y,c) ~ x1 + x2 + x3,  
               data=data, dist="weibull"))
```

Proportional hazard model

The base hazard can now have various different distributions, e.g. exponential.

In R:

```
summary(survreg(Surv(y,c) ~ x1 + x2 + x3,  
               data=data, dist="exponential"))
```

```
summary(survreg(Surv(y,c) ~ x1 + x2 + x3,  
               data=data, dist="weibull"))
```

Note that the coefficients enter multiplicatively, similar to count models. If $\beta_{x_1} = -0.16$, then the multiplicative effect is $e^{-0.16} = .85$, which an increase of x_1 by 1 leads to a 15% decrease in the hazard.

Cox proportional hazard model

The most commonly used proportional hazard model is **nonparametric**, i.e. there is no assumption made about the distribution of h_0 .

Cox proportional hazard model

The most commonly used proportional hazard model is **nonparametric**, i.e. there is no assumption made about the distribution of h_0 .

Using a nonparametric model leads to a slightly less efficient estimation, but a more generic one.

Time-varying independent variables

Two types of independent variables in survival analysis can be distinguished:

- Constant over time ($h_X(t) = h_0(t)e^{X\beta}$)
- Varying over time ($h_X(t) = h_0(t)e^{X(t)\beta}$)

Outline

- 1 Survival analysis
- 2 Application
- 3 Additional topics

Two data formats

Depending on whether there are time varying independent variables, a survival data set can be in two different formats.

	time	censored	x_1	x_2
Format 1:	10	1	3	1
	14	0	2	0
	13	1	5	1
	2	1	4	1

Two data formats

Depending on whether there are time varying independent variables, a survival data set can be in two different formats.

	start	end	event	censored	x_1	x_2
Format 2:	1	2	0	1	3	1
	2	3	0	0	2	0
	3	4	1	1	5	1
	1	2	0	1	3	1

Example: Recidivism

Open dataset *Rossi.txt*:

```
Rossi <- read.table(file.choose(), header=TRUE)
Rossi[1:5, 1:10]
```

Estimate a simple distribution:

```
library(survival)
summary(survreg(Surv(week, arrest) ~ 1, data=Rossi,
               dist="weibull"))
summary(survreg(Surv(week, arrest) ~
               fin + age + race + wexp + mar + paro + prio,
               data=Rossi, dist="weibull"))
```

Example: Recidivism

Estimate a Cox proportional hazard model:

```
Rossi.cph <- coxph(Surv(week, arrest) ~  
  fin + age + race + wexp + mar + paro + prio,  
  data=Rossi)
```

And a plot of the survival function:

```
plot(survfit(Rossi.cph), xlab="Weeks",  
  ylab="Proportion not rearrested")
```

Example: Recidivism

Producing a plot of predicted survival rates, with confidence intervals:

```
nd <- data.frame(cbind(c(0,1), mean(Rossi$age),
  mean(Rossi$race), mean(Rossi$wexp), mean(Rossi$mar),
  mean(Rossi$paro), mean(Rossi$prio)))
names(nd) <- c("fin","age","race","wexp",
  "mar","paro","prio")

plot(survfit(Rossi.cph, newdata=nd), xlab="Weeks",
  ylab="Proportion not rearrested",
  conf.int=TRUE, lty=c(1,2))
```

Example: Recidivism

Now to deal with time-dependent independent variables, first execute *aqm_2008_lecture_survival_fold.R*:

```
source(file.choose())
```

Then, fold the data and estimate the Cox proportional hazard model:

```
Rossi2 <- fold(Rossi, time="week", event="arrest",  
              cov=11:62, cov.names="employed", lag=1)
```

```
Rossi.cph2 <- coxph(Surv(start, stop, arrest.time) ~  
  fin + age + race + wexp + mar  
  + paro + prio + employed,  
  data=Rossi2)  
summary(Rossi.cph2)
```

Outline

- 1 Survival analysis
- 2 Application
- 3 Additional topics

Competing risks

The competing risks model is a survival model where there are multiple ways of failure or death, e.g. different causes of death.

Competing risks

The competing risks model is a survival model where there are multiple ways of failure or death, e.g. different causes of death.

In a competing risks model, right censoring can be included simply as another type of risk.

Frailty

In some scenarios, one wants to assume that the baseline hazard, $h_0(t)$, varies per individual, or per group of individuals.

Frailty

In some scenarios, one wants to assume that the baseline hazard, $h_0(t)$, varies per individual, or per group of individuals.

E.g. different types of companies might have different risks of bankruptcy.

Frailty

In some scenarios, one wants to assume that the baseline hazard, $h_0(t)$, varies per individual, or per group of individuals.

E.g. different types of companies might have different risks of bankruptcy.

A frailty is an extra parameter to a proportional hazard model that estimates this unit- or group-specific baseline hazard.

Frailty

In some scenarios, one wants to assume that the baseline hazard, $h_0(t)$, varies per individual, or per group of individuals.

E.g. different types of companies might have different risks of bankruptcy.

A frailty is an extra parameter to a proportional hazard model that estimates this unit- or group-specific baseline hazard.

$$h_X(t) = \alpha h_o(t) e^{X\beta} = h_o(t) e^{X\beta + \log(\alpha)},$$

with typically $\log(\alpha_i) \sim N(0, \sigma^2)$.