Statistics for Life Sciences - Survival Analysis

Nathan Harmston

2022-03-29

from an **exposure** to an **event**

Time-to-Event Analysis

In survival analysis, we are interested not only in outcomes, but also in the time it takes for them to occur. This time variable can be referred to as failure time, survival time or event time.

Instances where you would use survival analysis in biological research

- Time from exposure to onset of symptoms
- Time from cancer treatment until death
- Time until seizure freedom after taking an anti-epileptic drug
- lifespan of flies on distinct sugar diets

from an **exposure** to an **event**

Time-to-Event Analysis

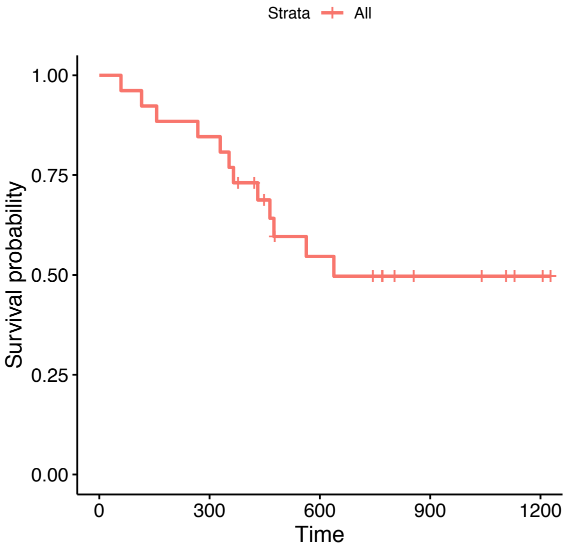
In survival analysis, we are interested not only in outcomes, but also in the time it takes for them to occur. This time variable can be referred to as failure time, survival time or event time.

Instances where you would use survival analysis in biological research

- Time from exposure to onset of symptoms
- Time from cancer treatment until death
- Time until seizure freedom after taking an anti-epileptic drug
- lifespan of flies on distinct sugar diets

marriage .... divorce

Kaplan-Meier curves



```
sd <- survdiff(Surv(futime, fustat) ~ resid.ds,
         data = ovarian)
sd
1 - pchisq(sd$chisq, length(sd$n) - 1)
```

$$\chi^2 = \sum \frac{(o-e)^2}{e} \tag{1}$$

where do the expected really come from in this case ?

- order by time point
- at each time point calculate the number of expected deaths in group a and in group b
- sum them up

log-rank test

$$\chi^2 = \frac{(3 - 6.26)^2}{6.26} + \frac{(9 - 5.74)^2}{5.74} \tag{2}$$

$$\text{e} = \text{number at risk in group 1 at time t} * \frac{\text{total number of events at time t}}{\text{total number at risk}} \tag{3}$$

Survival function

where the $S(t)$ gives the probability that a subject will survive past time t

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

time	status
2	1
3	0
6	1
6	1
7	1
10	0
15	1
15	1

Instantaneous death or failure rate

The hazard function $h(t)$ is the instantaneous risk that the event of interest happens, within a very narrow time frame.

$$h(t) = Pr(t < T \leq t + \Delta t | T > t)$$

Instantaneous death or failure rate

The hazard function $h(t)$ is the instantaneous risk that the event of interest happens, within a very narrow time frame.

$$h(t) = Pr(t < T \leq t + \Delta t | T > t)$$

the probability of dying in the few Δ given you are alive right now

Hazard function

what are corresponding values of the hazard function for our example dataset?

time	status
2	1
3	0
6	1
6	1
7	1
10	0
15	1
15	1

Hazard Ratio

$$HR = \frac{HAZ(X = 1)}{HAZ(X = 0)}$$

exposure versus non-exposure

A HR < 1 indicates reduced hazard of death whereas a HR > 1 indicates an increased hazard of death.

if HR $= 3$ - then your risk of death is 3x if exposed compared to non-exposed

rate of decrease of survival curve

The HR is interpreted as the instantaneous rate of occurrence of death (the event) of interest in those who are still at risk of death (the event).

Cox Proportional Hazard Models

We may want to quantify an effect size for a single variable, or include more than one variable into a regression model to account for the effects of multiple variables

regress, not the time-to-events themselves, but the failure or hazard rates onto explanatory variables

$$h(t|Z) = h_0(t)e^{\beta Z}$$

Cox Proportional Hazard Models

$$h(t|Z) = h_0(t)e^{\beta Z}$$

$$\frac{h(t)}{h_0} = e^{\beta Z}$$

$$log(\frac{h(t)}{h_0}) = \beta Z$$

$$log(\frac{h(t)}{h_0}) = \beta_1 Z_1 + \beta_2 Z_2 + \beta_3 Z_3 + \beta_3 Z_3 + \beta_4 Z_4$$

$$HR = \frac{HAZ(X=1)}{HAZ(X=0)}$$

hazards can change over time but the hazard ratio stays constant

how do we check this assumption?

?cox.zph

Ovarian Cancer

This dataset from a cohort of ovarian cancer patients. It contains clinical information, including age, treatment group, presence of residual disease, performance, blood pressure*, cholesterol levels*, **if the subjects were censored or not and the time subjects were tracked until they either died or were lost to follow-up.** The patients were followed up for ~ 3.5 years after treatment.

Age - age at treatment
Resid Disease - was there residual disease after treatment
Rx - which drug were they put on, A or B
ECOG - quality of life
BP - blood pressure at start of treatment
Chol - cholesterol levels at start of treatment
Death - 1 if the subject died or 0 if they were censored
Time - time until death or censoring

Ovarian Cancer

	Age	Resid Dis	Rx	ECOG	BP	Chol	**Death**	**Time**
1	72.33	yes	A	good	117.83	13.58	**1**	**59**
2	74.49	yes	A	good	114.00	7.78	**1**	**115**
3	66.47	yes	A	bad	117.55	10.95	**1**	**156**
4	74.50	yes	A	bad	113.50	22.50	**1**	**268**
5	43.14	yes	A	good	139.19	22.11	**1**	**329**
6	63.22	no	B	bad	124.80	8.46	**1**	**353**
7	64.42	yes	B	good	118.09	23.19	**1**	**365**
8	58.31	no	B	good	130.09	26.51	**0**	**377**
.	.	.	.	.	.	.	**.**	**.**
.	.	.	.	.	.	.	**.**	**.**
25	44.21	yes	B	good	138.34	26.25	**0**	**1206**
26	59.59	no	B	bad	129.18	22.93	**0**	**1227**

Cox regression model

```
fit.coxph <- coxph(surv_object ~ rx + resid.ds +
                   age_group + ecog.ps,  data = ovarian)
fit.coxph
```

Cox regression model

```
> fit.coxph
Call:
coxph(formula = surv_object ~ rx + resid.ds + age_group + ecog.ps,
    data = ovarian)

                 coef exp(coef) se(coef)      z      p
rxB           -1.3814    0.2512   0.6448 -2.142 0.0322
resid.dsyes    1.4470    4.2503   0.7292  1.984 0.0472
age_groupyoung -2.2013    0.1107   1.1069 -1.989 0.0467
ecog.psbad     0.5859    1.7966   0.6329  0.926 0.3546

Likelihood ratio test=12.19  on 4 df, p=0.01596
n= 26, number of events= 12
```

The quantity of interest from a Cox regression model is a hazard ratio (HR).

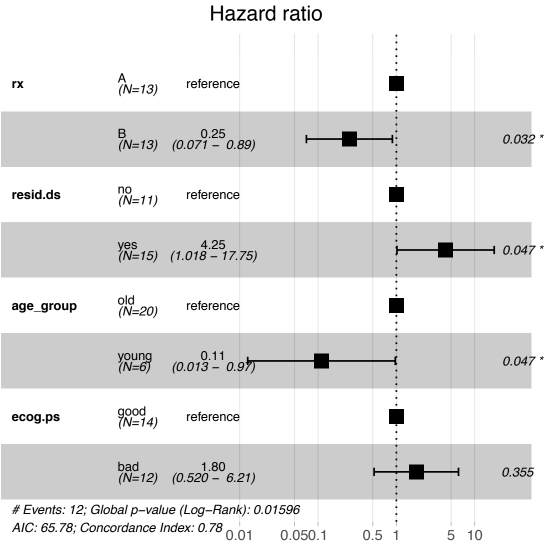
The HR represents the ratio of hazards between two groups at any particular point in time.

If you have a regression parameter (from column estimate from coxph) then HR = $\exp(\beta)$.

So HR = 4 implies that around 4 times as many patients with residual disease are dying compared to patients without residual disease, at any given time.

Hazard ratio - forest plot

```
ggforest(fit.coxph, data = ovarian)
```



Case study - Lung Cancer

This dataset from a cohort of lung cancer patients from the North Central Cancer Treatment Group. It contains clinical information, the time subjects were tracked until they either died or were lost to follow-up and drug treatment.

inst: Institution code
time: Survival time in days
status: censoring status 1=censored, 2=dead
age: Age in years
sex: Male=1 Female=2
ph.ecog: ECOG performance score (0=good 5=dead)
ph.karno: Karnofsky performance score (bad=0-good=100) rated by physician
pat.karno: Karnofsky performance score as rated by patient
meal.cal: Calories consumed at meals
wt.loss: Weight loss in last six months

Case study - Lung Cancer

	inst	time	status	age	sex	ph.ecog	ph.karno	pat.karn
1	3.00	306.00	2.00	74.00	1.00	1.00	90.00	100.0
2	3.00	455.00	2.00	68.00	1.00	0.00	90.00	90.0
3	3.00	1010.00	1.00	56.00	1.00	0.00	90.00	90.0
4	5.00	210.00	2.00	57.00	1.00	1.00	90.00	60.0
5	1.00	883.00	2.00	60.00	1.00	0.00	100.00	90.0
6	12.00	1022.00	1.00	74.00	1.00	1.00	50.00	80.0

Using this dataset, I want you to come up with

- Two survival analysis models

and to describe the potential implications/assumptions of these analyses