

# 4. Parametric estimation and Proportional Hazard Models

---

CHRIS SUTTON, OCTOBER 2023

# Parametric estimation

---

# a different approach

---

Kaplan-Meier is an example of non-parametric estimation

We now turn to the alternative, parametric estimation of the survival function

Here we will assume  $S(t)$  has some functional form

- this means we can express  $S(t)$  and the hazard  $\mu_t$  in terms of parameters of that function
- we can then seek to estimate these parameters by maximum likelihood techniques
- we have already met some simple functional forms of  $S(t)$  when we introduced exponential; Weibull; Gompertz; Makeham in week 2

# Maximum Likelihood Estimation

---

# What is a MLE?

---

The maximum likelihood estimator  $\hat{\theta}$  for a parameter  $\theta$ , is the estimate which maximises the probability of obtaining the sample we have actually observed

The maximum likelihood estimator is the parameter estimate that maximises the “likelihood function” which is the joint probability function [discrete distribution] or joint pdf [continuous] of the observed sample

# Binomial example

---

n trials gives observations  $y_1, y_2, \dots, y_n$

- $y_i = 1$  if the  $i^{\text{th}}$  trial a success;  $y_i = 0$  otherwise

we look for the MLE of the probability of success “w”

likelihood function is,  $L(w) = L(y_1, y_2, \dots, y_n \mid w) = w^y (1 - w)^{n-y}$

- where  $y = \sum y_i$
- the MLE is the w which maximises  $L(w)$
- we do this by taking derivative of  $L(w)$  with respect to w; setting to zero and solving for w
- as  $L(w)$  is a product of functions it is much easier to find the derivative of  $\log[L(w)]$  given  $L(w)$  and  $\log[L(w)]$  are maximised at the same value of w

# Binomial example continued

---

$$\begin{aligned}\log[L(w)] &= \log[w^y (1-w)^{n-y}] \\ &= y \log(w) + (n-y) \log(1-w)\end{aligned}$$

and

$$\frac{d}{dw} \log[L(w)] = y \left[ \frac{1}{w} \right] + (n-y) \left[ \frac{-1}{1-w} \right] = \frac{y}{w} - \frac{n-y}{1-w}$$

the log likelihood (and likelihood) are maximised at  $\hat{w}$  where

$$\frac{y}{\hat{w}} - \frac{n-y}{1-\hat{w}} = 0 \quad \text{which solves to give } \hat{w} = \frac{y}{n} \quad (\text{the fraction of success observed in the trials, which is intuitive})$$

[we should now take 2<sup>nd</sup> derivative and check <0 to make sure we have found max not min]

# MLE in Survival Models

---



# MLE with the exponential model

---

We will use the exponential hazard model as a working example of a parametric approach to estimating the lifetime distribution

Recall the exponential model assumes the hazard (or force of mortality) is a constant  $\mu$

This gives us  $S_x(t) = \exp(-\mu t)$

But  $\mu$  here is an unknown. Today we are seeking the estimate of  $\mu$  which is most likely given our observations. The maximum likelihood estimate is denoted  $\hat{\mu}$

# observations

---

n lives observed from exact age  $x$  until one of:

- a) death
- b) withdrawal from the investigation during the year
- c)  $x+1^{\text{th}}$  birthday

note that b) and c) are forms of censoring

we will consider first category a) and then b) + c)

# deaths

---

If:

- there are  $k$  deaths at durations  $t_1, t_2, \dots, t_k$  and
- $f(t)$  is the probability density function of lifetime  $T$

Then

- the probability that life 1 actually dies at duration  $t_1$  is  $f(t_1)$
- the prob that life 1 dies at duration  $t_1$  and life 2 dies at duration  $t_2$  is  $f(t_1) \cdot f(t_2)$
- for all  $k$  deaths the probability is

$$\prod_{\text{all deaths}} f(t_i)$$

# censored

---

if the first censored life was censored at duration  $t_{k+1}$

all that we know is that life survived to at least  $t_{k+1}$

- the probability of this is  $S(t_{k+1})$

so the probability of observing all the data we have for censored lives is

$$\prod_{\text{all censored}} S(t_i)$$

# Likelihood function

---

putting together the deaths and the censored lives, the probability of observing all the data that we observed is:

$$L = \prod_{\text{all deaths}} f(t_i) \cdot \prod_{\text{all censored}} S(t_i)$$

this product is the ‘likelihood’ of the observed data

we seek a MLE  $\hat{\mu}$  of the exponential model hazard parameter  $\mu$  which will maximise this likelihood product  $L$

# finding the MLE $\hat{\mu}$

---

find the likelihood function  $L$  in terms of  $\mu$

differentiate  $L$  [or, more often  $\log(L)$ ] with respect to  $\mu$

set to zero and solve for  $\hat{\mu}$

# Demonstration

---

See Demonstration07  
pdf file on QM Plus

# $\hat{\mu}$ in the exponential model

---

therefore

$$\hat{\mu} = \frac{\sum \delta_i}{\sum t_i} = \frac{\text{total number of deaths}}{\text{total time lives in study exposed to risk of death}}$$

This is an example of parametric estimation with one parameter ( $\mu$  here)

If we had more than one parameter we would need to differentiate  $L$  with respect to each one and then solve simultaneous equations (which in most cases would mean iterative rather than analytical methods)



# Chaining probabilities together

---

# chaining

---

In practice with human mortality we find no one survival function is accurate over all ages

It is better therefore to sub-divide ages and 'chain together' several different functions

This is simple to do with our exponential model example

- if the MLE for the force of mortality for single year from age  $x$  to  $x+1$  is  $\hat{\mu}_x$
- then the survival function for that year is  $\hat{S}_x(1) = \exp(-\hat{\mu}_x)$
- then if we estimate the force of mortality for the next year to be  $\hat{\mu}_{x+1}$
- the probability a person alive at age  $x$  is still alive at age  $x+2$  is

$$\hat{S}_x(2) = \exp(-\hat{\mu}_x) \cdot \exp(-\hat{\mu}_{x+1}) = \exp\{-(\hat{\mu}_x + \hat{\mu}_{x+1})\}$$

# general exponential model chain

---

chaining together  $m$  years from age  $x$  with a MLE of the force of mortality in each successive year:

$$\hat{S}_x(m) = \exp \left[ - \sum_{j=0}^{m-1} \hat{\mu}_{x+j} \right]$$



# Proportional Hazard Models

---

CHRIS SUTTON

OCTOBER 2023

# Topic outline

---

1

- Covariate data

2

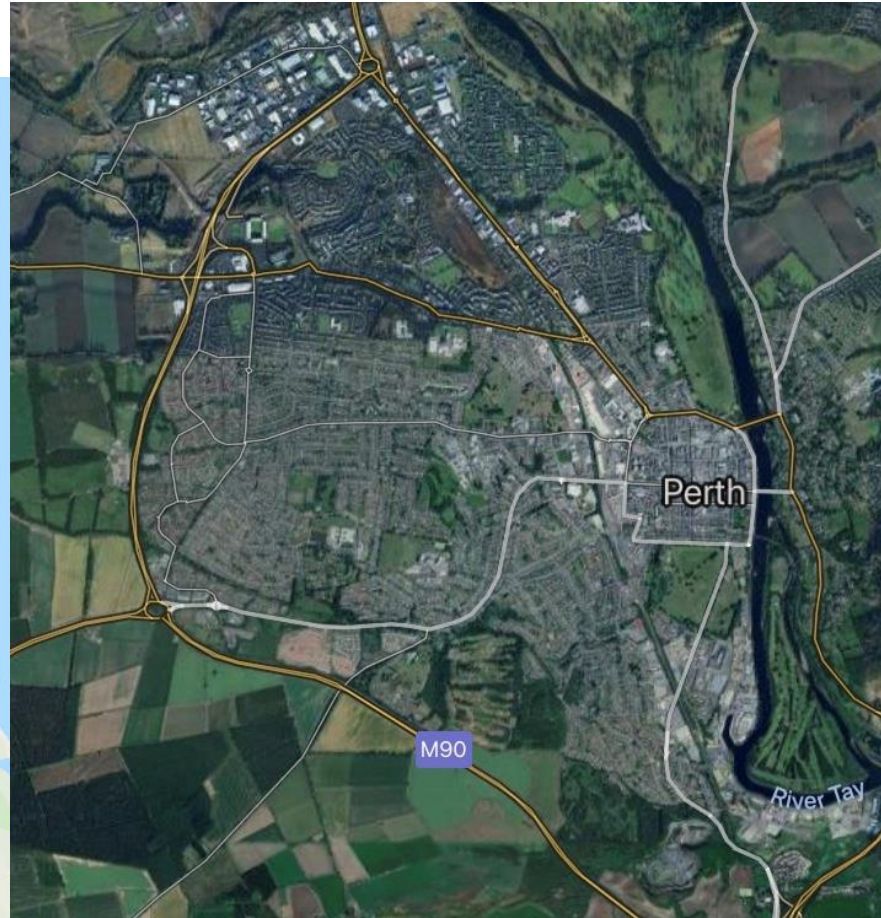
- Proportional Hazard (PH) models

3

- The Cox PH model

4

- Model fitting criteria



If you were asked to carry out an investigation comparing mortality in Perth, Scotland and Perth, Australia what data (in addition to observed deaths) would you collect on the population of the 2 cities?

# Covariate data

---



# Covariates

---

so far the models and estimators we have looked at in this course have used only age and duration data ( $x, t$ )

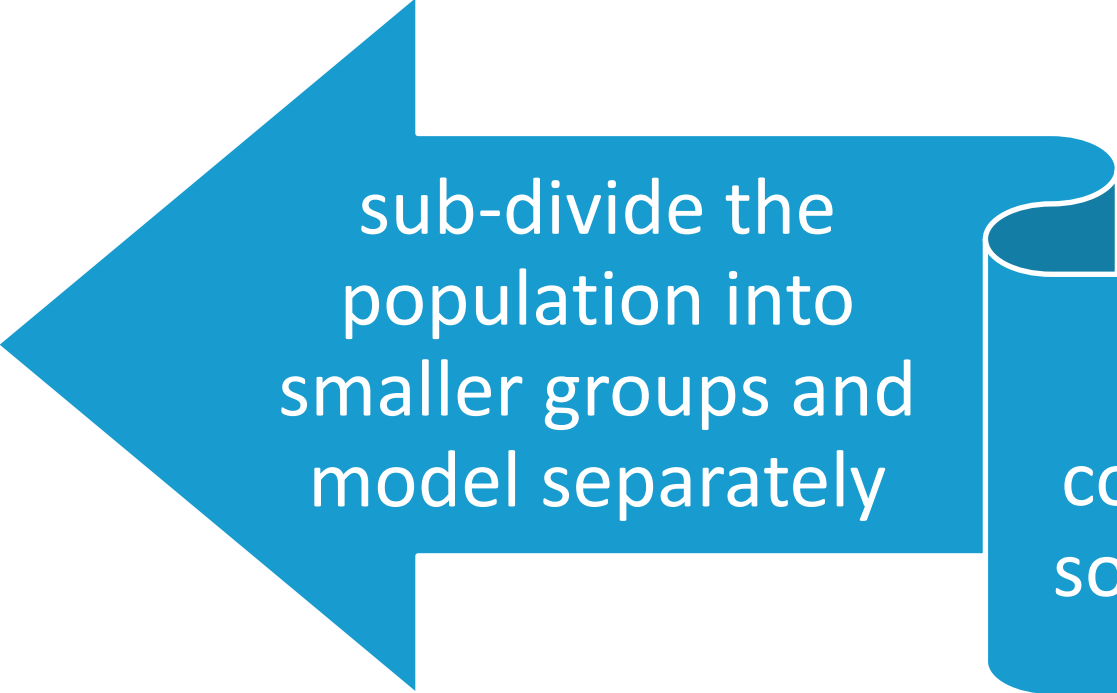
in practice more data would usually be recorded for each life in a study which might be valuable for modelling purposes

Age	Male / Female	Smoker / non-smoker	Type of treatment	Symptom severity	Postcode	Time since last medical
-----	------------------	------------------------	----------------------	---------------------	----------	----------------------------

this data is called **covariate data**

# 2 ways to deal with covariate data

---



sub-divide the population into smaller groups and model separately



model the effects of covariates directly using some 'regression model'

# covariate notation

---

there are  $p$  covariate data measures obtained for each life

$z_i$  is a  $1 \times p$  vector of covariates for the  $i^{\text{th}}$  life

$$z_i = ( X_{i1}, X_{i2}, \dots, X_{ip} )$$

the covariates can be collected in one of three ways

Covariate measure	Example time since last medical
Raw numerical value	Actual time in months
0 or 1 value assigned	1 if within last year, 0 otherwise
Score on some other scale e.g. 1 to 5 (qualitative)	1 if 0-3 months; 2 if 3-6 months; 3 if 6-12 months; 4 if 12-24 months; 5 if > 2 years

# example covariates

---

Life	Male or Female (F=1; M=0)	Weekly alcohol consumption (units)	Time since last medical (scored 1-5)	Prior history of heart disease (yes=1; no=0)
1	1	6	5	0
2	1	0	3	0
3	0	26	2	1
4	1	12	2	0
5	0	9	5	0

# Proportional Hazard models

---

# PH models

---

The most commonly used regression models in survival analysis

- can be built using non-parametric or parametric approaches

Instead of the force of mortality  $\mu_{x+t}$  used in our simpler models, we introduce the **hazard function**  $\lambda_i$  for the  $i^{\text{th}}$  life where the hazard is now a function of both duration  $t$  and the covariate data vector  $z_i$

In a Proportional Hazard model,

$$\lambda_i(t, z_i) = \lambda_0(t) \cdot g(z_i)$$

where  $\lambda_0(t)$  is a function of duration only

and  $g(z_i)$  is a function of the covariate vector only

# PH model simplifying assumption

---

$$\lambda_i(t, z_i) = \lambda_0(t) \cdot g(z_i)$$

$\lambda_0(t)$  a function of  $t$  only is the “baseline hazard”, it is the hazard for an individual with covariate vector of zero

$g(z_i)$  is a function of covariate data only

This is much simpler to model than a single function that varies with both covariate data and duration

# Parametric Proportional Hazard model

---



# parametric PH model

---

we can construct PH models on a non-parametric or parametric basis

in general in a parametric model we assume the lifetime distribution follows a certain functional form

in a parametric PH model we assume the hazard function follows one of the types of parametric survival models e.g. exponential; Gompertz; Makeham (or others) and we bring the covariate vector into the model parameters

# parametric PH using Gompertz

---

## Example:

we know Gompertz as  $\mu_x = Bc^x$  for some parameters  $B$  and  $c$

in terms of a hazard function rather than a force of mortality this translates to

$$\lambda(t) = Bc^t$$

Now in a PH model we can let parameter  $B$  depend on the covariates

if  $z_i$  is our  $1 \times p$  covariate vector and  $z_i^T$  is the transpose of that vector (so  $p \times 1$ )

then we can set Gompertz parameter  $B$  to be  $B = \exp(\beta \cdot z_i^T)$

where  $\beta$  is a  $1 \times p$  vector of regression coefficients  $(\beta_1, \beta_2, \dots, \beta_p)$

# PH with Gompertz e.g. cont'd

---

$$\lambda_i(t, z_i) = c^t \exp(\beta \cdot z_i^T)$$

where

- the baseline hazard is  $c^t$
- $\exp(\beta \cdot z_i^T)$  is the [assumed] effect of the covariates

the log-hazard is linear and separates the baseline hazard term

$$\log[\lambda_i(t, z_i)] = t \log(c) + \beta \cdot z_i^T \text{ which can be very convenient to work with}$$

however the usefulness of this model will depend entirely on how effectively we can estimate the regression coefficients  $\beta$  [the  $1 \times p$  vector  $(\beta_1, \beta_2, \dots, \beta_p)$ ]

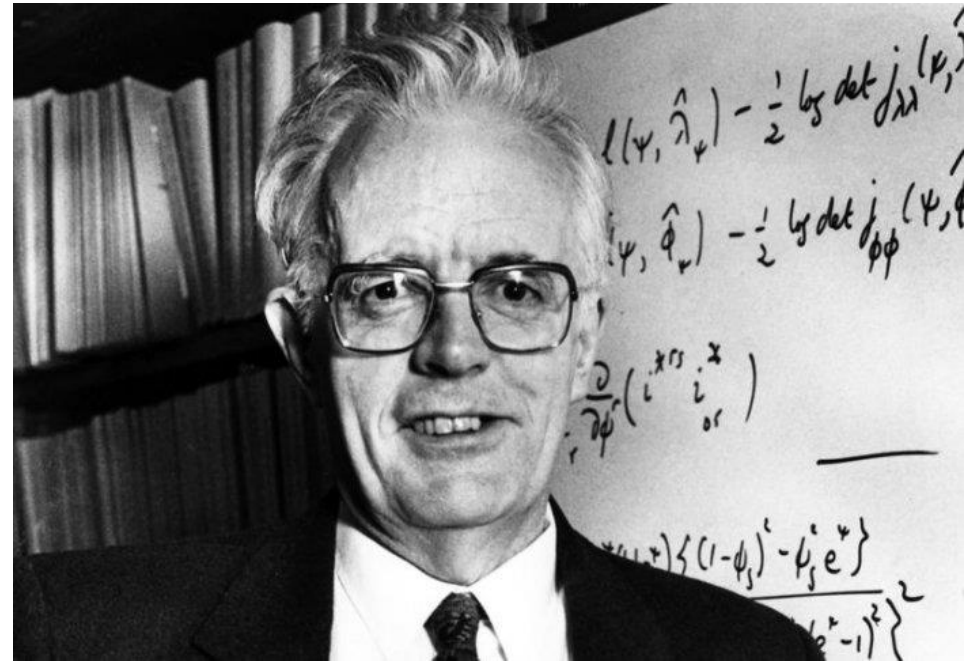
# The Cox PH model

---

# introduction to Cox model

Cox, D.R. (1972) 'Regression Models and Life-Tables', *Journal of the Royal Statistical Society* (series B, methodological ) vol.34(2) pp.187-220

- Imperial College London
- won the first International Statistics Prize in 2016 (awarded by the American Statistical Association, see [statprize.org](http://statprize.org)) for this survival models work



Professor Sir David Cox, 1980. Source: General Motors Cancer Research Foundation, National Cancer Institute.

# what and why

---

situations where we do not need to know the precise rate of mortality but instead are interested in relative levels of mortality between different individuals

we assume each individual's mortality is proportional to some general function (the baseline hazard)

- we do not worry about the shape of this baseline hazard
- instead we focus on the constant of proportionality for each individual which will depend on the covariates
- this is a widely used survival model

# Cox PH model

---

Hazard is in the form

$$\lambda_i(t, z_i) = \lambda_0(t) \exp(\beta \cdot z_i^\top)$$

- so the general shape of the hazard function depends on the baseline hazard  $\lambda_0(t)$
- the differences between individuals are given by  $\beta \cdot z_i^\top$
- if we are conducting a study (e.g. a medical trial) where we are more interested in the effect of covariates than in the shape of the hazard, this means we can ignore the baseline hazard  $\lambda_0(t)$  and look to estimate the regression coefficients  $\beta$  irrespective of  $\lambda_0(t)$
- this is called a “semi-parametric” approach which is widely used in statistical survival models
- the Cox model uses the method of “partial likelihood” to estimate the coefficients  $\beta$  but not the baseline hazard
- partial likelihood statistics behave in similar way to the more usual maximum likelihood

# Partial likelihood in Cox

---

assume deaths are observed at times  $t_1, t_2, \dots, t_k$  with just one death at each  $t_j$

let  $R(t_j)$  be the set of lives at risk of death at time  $t_j$  (just prior to the  $j^{\text{th}}$  death)

the partial likelihood calculation depends only on the order in which deaths are observed

the probability that life 1 [out of the set  $R(t_1)$ ] is the life that dies at  $t_1$  (conditional on one death being observed at that time) is

$$\frac{\lambda_0(t) \exp(\beta \cdot z_1^T)}{\sum_{i \in R(t_1)} \lambda_0(t) \exp(\beta \cdot z_i^T)}$$

← the baseline hazard  $\lambda_0(t)$  will cancel top and bottom here



# Partial likelihood in Cox (cont'd)

---

repeating this calculation for all observed deaths at  $t_1, t_2, \dots, t_k$  the probability that life 1 out of set  $R(t_1)$  dies at time  $t_1$  and life 2 out of set  $R(t_2)$  dies at time  $t_2$  and .... life  $k$  out of set  $R(t_k)$  dies at time  $t_k$  is given by the product of these probabilities.

This is the **partial likelihood function** which here is a function of parameters  $\beta$  (the regression coefficients)

$$L(\beta) = \prod_{j=1}^k \frac{\exp(\beta \cdot z_j^T)}{\sum_{i \in R(t_j)} \exp(\beta \cdot z_i^T)}$$

# Partial likelihood in Cox

---

this  $L(\beta)$  is a partial likelihood function because it only uses the order in which deaths are observed and the rest of the observed data is discarded

To find the likelihood estimates for the regression coefficients  $\beta$  we would need to differentiate with respect to each of the  $p$  coefficients that make up  $\beta$ , set to zero and solve for vector  $\hat{\beta}$  our likelihood estimate of vector  $\beta$ .

This vector of derivatives which we set to zero is the **efficient score function**  $u(\beta)$

$$u(\beta) = \left[ \frac{d \log L(\beta)}{d\beta_1}, \dots, \frac{d \log L(\beta)}{d\beta_p} \right]$$

then  $\hat{\beta}$  found by solving  $u(\hat{\beta}) = 0$

in practice we will not be able to do this algebraically but will need a computer package

# Breslow's approximation

---

The Cox PH model assumes there is one death at time  $t_j$ . If instead there are  $d_j > 1$  deaths at time  $t_j$ , the modelling becomes much more complex because all of the possible combinations of the  $d_j$  deaths out of  $R(t_j)$  need to be included in the likelihood function.

In this scenario, "Breslow's approximation" is sometimes used:

$$L(\beta) \approx \prod_{j=1}^k \frac{\exp(\beta \cdot s_j^T)}{\left[ \sum_{i \in R(t_j)} \exp(\beta \cdot z_i^T) \right]^{d_j}}$$

← where  $s_j$  is the sum of the  $z$  covariate vectors for the  $d_j$  lives observed to die at time  $t_j$

# Model fitting criteria

---

# assessing covariates

---

In PH models (including Cox) we need criteria for assessing the effects of the different covariates

the **likelihood ratio statistic** gives one method for doing this

model 1 has  $p$  covariates

model 2 has additional  $q$  covariates (so  $p+q$  in total)

$\log L_p$  = the maximised log-likelihood of model 1

$\log L_{p+q}$  = the maximised log-likelihood of model 2

then

likelihood ratio statistic =  $-2(\log L_p - \log L_{p+q})$

# likelihood ratio statistic comments

---

generally the likelihood ratio statistic will use the full likelihood function (the one used to derive MLEs) but for the Cox model it is okay to use it with partial likelihoods

this likelihood ratio statistic has an asymptotic [that is it approaches as a limit]  $\chi^2$  distribution on  $q$  degrees of freedom under the hypothesis that the additional  $q$  covariates have no effect when the first  $p$  covariates are there

further tests of interactions between different covariates will not be covered in this module

# 2 model building strategies

---



start with null model that has no covariates then add new covariates one at a time and evaluate with likelihood ratio statistic



begin with full model that has all the possible covariates and use likelihood ratio statistics to eliminate covariates that have no statistically significant affect

