SPA 2403 SURVIVAL ANALYSIS

Cox-PH Models

Dr. Mutua Kilai

Kirinyaga University

Introduction

The Cox PH model is written in terms of the hazard model formula.

This model gives an expression for the hazard function at time t for an individual with a given specification of a set of explanatory variables denoted by X

The model is given as

$$h(t, X) = h_0(t)e^{\sum_{i=1}^{p}\beta_i X_i}$$

The term $h_0(t)$ is the baseline hazard function.

The explanatory variables are said to time-independent.

An important feature of this formula,which concerns the proportional hazards (PH) assumption, is that the baseline hazard is a function of t, but does not involve the X's.

The Cox model formula has the property that if all the X's are equal to zero, the formula reduces to the baseline hazard function. That is, the exponential part of the formula becomes e to power zero, which is 1.

A time-independent variable is defined to be any variable whose value for a given individual does not change over time. Example is gender.

The baseline hazard $h_0(t)$ is an unspecified function, this property makes the Cox model a **semiparamtric** model.

The cumulative hazard is

$$\begin{aligned}\Lambda(t|X) &= \int_0^t \lambda(t|X)dt \\ &= \int_0^t \lambda_0(t)\exp(X'\beta)dt \\ &= \{\int_0^t \lambda_0(t)dt\}\exp(X'\beta) \\ &= \Lambda_0(t)\exp(X'\beta)\end{aligned} \tag{1}$$

Here $\Lambda_0(t)$ is called the baseline cumulative hazard function.

Let's derive the survival function in this scenario

$$S(t|X) = \exp\{-\Lambda(t|X)\} = \exp\{-\Lambda_0(t)\exp(X'\beta)\}$$

The density function is given by:

$$\begin{aligned} f(t|X) &= -\frac{d}{dt}S(t|X) \\ &= -\frac{d}{dt}\exp\{-\Lambda_0(t)\exp(X'\beta)\} \\ &= \exp\{-\Lambda_0(t)\exp(X',\beta)\}\lambda_0(t)\exp(X'\beta) \\ &= S(t|x)\lambda(t|X) \end{aligned} \tag{2}$$

Cox-PhH Estimation

For the observed data (V_i, Δ_i, X_i) $i = 1, ..., n$ the likelihood for the Cox PH model is

$$\begin{aligned} L(\beta) &= \prod_{i=1}^{n} f^{\Delta_i}(V_i|X_i)\{S(V_i|X_i)\}^{1-\Delta_i} \\ &= \prod_{i=1}^{n}\{\lambda(V_i|X_i)S(V_i|X_i)\}^{\Delta_i}\{S(V_i|X_i)\}^{1-\Delta_i} \\ &= \prod_{i=1}^{n}\{\lambda_0(V_i\exp(X_i'\beta)\}^{\Delta_i}\exp\{-\Lambda_0(V_i\exp(X_i')\} \end{aligned} \tag{3}$$

To estimate β by maximizing $L(\beta)$ one may specify a parametric form for the function $\lambda_0(.)$. Once the functional form for the λ_0.

Once the functional form of λ_0 is specified, the model becomes a parametric model.

In a semi-parametric model Cox PH λ_0 is left unspecified.

Parametric form for $\lambda_0(.)$

If $\lambda_0(t) = c_0$ a constant, we obtain the exponential model.

If $\lambda_0(t) = c_0t^{c_1}$ a polynomial in t we obtain the Weibull model.

Cox PH Model λ_0 is unspecified estimation

For the semiparametric model $(\lambda(t|X)) = \lambda_0(t)\exp(X'\beta)$ Cox proposed to estimate β by maximizing the partial likelihood function

$$L_p(\beta) = \prod_{i=1}^{n}\{\frac{\exp(X_i')}{\sum_{j\in R(V_i)}\exp(X_j'\beta)}\}^{\Delta_i}$$

$R(V_i)$ is the risk set at time V_i comprised of all individuals with survival or censoring times $\geq V_i$

Cox PH Model Estimation

To maximize $L_p(\beta)$ we first log transform $L_p(\beta)$

$$l_p(\beta) = \sum_{i=1}^{n}\Delta_i\Big[X_i'\beta - \log\{\sum_{j\in R(V_i)}\exp(X_j'\beta)\}\Big]$$

then differentiate

$$\frac{\partial}{\partial\beta} l_p(\beta) = \sum_{i=1}^{n} \Delta_i \Big\{ X_i - \frac{\sum_{j \in R(V_i) X_j \exp(X'_j\beta)}}{\sum_{j \in R(V_i) \exp(X'_j\beta)}} \Big\}$$

and we can solve $\frac{\partial}{\partial\beta} l_p \beta = 0$ by numerical methods to obtain $\hat{\beta}$

The estimator of the baseline hazard is

$$\hat{\lambda_0}(t) = \frac{\Delta_k}{\sum_{j \in R(V_k) \exp(X'_j\hat{\beta})}} \quad if \;\; t = V_k$$

The estimator of the cumulative baseline hazard is

$$\hat{\Lambda_0}(t) = \int_0^t \hat{\lambda_0}(u) du = \sum_{V_k \leq t} \frac{\Delta_k}{\sum_{j \in R(V_k)} \exp(X'_j\beta)}$$

The estimator of the survival function at time τ is

$$\hat{S}(\tau|X) = \exp\{-\hat{\Lambda_0}(\tau)\exp(X^T\hat{\beta})\}$$

Cox PH Model Standard Errors

We can estimate $Var(\hat{\beta})$ by $I^{-1}(\hat{\beta})$ where

$$I(\beta) = -\frac{\partial^2 l_p(\beta)}{\partial\beta\partial\beta'}$$

is called the information matrix and $I(\hat{\beta})$ is obtained by plugging $\hat{\beta}$ in for β.

Standard errors for $\hat{\beta}$ are then the square root of the diagonal elements of $I^{-1}(\hat{\beta})$

Wald Test

Suppose we are interested in testing the $j-th$ component of the β vector.

Suppose that $H_0 : \beta_j = \beta_j^*$ versus $H_1 : \beta_j \neq \beta_j^*$ Then the test statistic is

$$T = \frac{\hat{\beta}_j - \beta_j^*}{se(\hat{\beta}_j)}$$

which approximately follows $N(0,1)$ under the null hypothesis.

Assumptions of the Cox-PH Model

- Independent Observations: This assumption means that there is no relationship between the subjects in your data set and that information about one subject's survival does not in any way inform the estimated survival of any other subject.

- Non-informative or Independent censoring: This assumption is satisfied when there is no relationship between the probability of censoring and the event of interest. Violations of this assumption invalidate the estimates and p-values of the CPH model.

- The survival curves for two different strata of a risk factor must have hazard functions that are proportional over time: This assumption is satisfied when the change in hazard from one category to the next does not depend on time. This why the model is called the proportional hazards model.

Example

Consider the Veteran Lung Cancer data given in the `survival` package in R

```r
library(survival)
data(lung)
head(lung)
```

```
##   inst time status age sex ph.ecog ph.karno pat.karno meal.cal wt.loss
## 1    3  306      2  74   1       1       90       100     1175      NA
## 2    3  455      2  68   1       0       90        90     1225      15
## 3    3 1010      1  56   1       0       90        90       NA      15
## 4    5  210      2  57   1       1       90        60     1150      11
## 5    1  883      2  60   1       0      100        90       NA       0
## 6   12 1022      1  74   1       1       50        80      513       0
```

The covariates in the data are:

- sex $z_i = 1,\ denotes\ Male\ and\ 2\ denotes\ Female$
- age x_i

The model thus becomes

$$\lambda_i(t) = h_0(t)\exp\{\beta_1 x_i + \beta_2 z_i\}$$

The class of lives to which the baseline hazard function applies is the individuals whose age is 0, gender male,

We fit the Cox-PH model as

```r
out=coxph(Surv(time, status)~age+as.factor(sex),data=lung)
summary(out)
```

```
## Call:
## coxph(formula = Surv(time, status) ~ age + as.factor(sex), data = lung)
##
##   n= 228, number of events= 165
##
##                      coef exp(coef)  se(coef)      z Pr(>|z|)
## age              0.017045  1.017191  0.009223  1.848  0.06459 .
## as.factor(sex)2 -0.513219  0.598566  0.167458 -3.065  0.00218 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                 exp(coef) exp(-coef) lower .95 upper .95
## age                1.0172     0.9831    0.9990    1.0357
## as.factor(sex)2    0.5986     1.6707    0.4311    0.8311
##
## Concordance= 0.603  (se = 0.025 )
## Likelihood ratio test= 14.12  on 2 df,   p=9e-04
## Wald test            = 13.47  on 2 df,   p=0.001
## Score (logrank) test = 13.72  on 2 df,   p=0.001
```

The Cox regression output can be interpreted as:

- *Statistical Significance*: Column marked z gives the wald statistic. We have its p-value and from the output the p-value of sex being female is significant at 5% level.

- *Hazard Ratios*: The exponentiated coefficients gives the effect size of the covariates. For example being a female (sex = 2) reduces the hazard by a factor of 0.51.

- *Global Statistical significance of the model*: Finally, the output gives p-values for three alternative tests for overall significance of the model: The likelihood-ratio test, Wald test, and score logrank statistics. They should be less than 0.05

The fitted model becomes

$$\lambda_i(t) = h_0(t) \exp\{0.017x_i - 0.513z_i\}$$

What does the model tell you about the relative risk of a male aged 25 at entry compared to a female aged 40 at entry?

Here we compute the ratios as follows

$$\lambda_i(t) = \frac{h_0(t) \exp\{0.017 \times 25 - 0.513 \times 1\}}{h_0(t) \exp\{0.017 \times 40 - 0.513 \times 2\}} = \frac{\exp(-0.088)}{\exp(0.663)} = 0.471$$

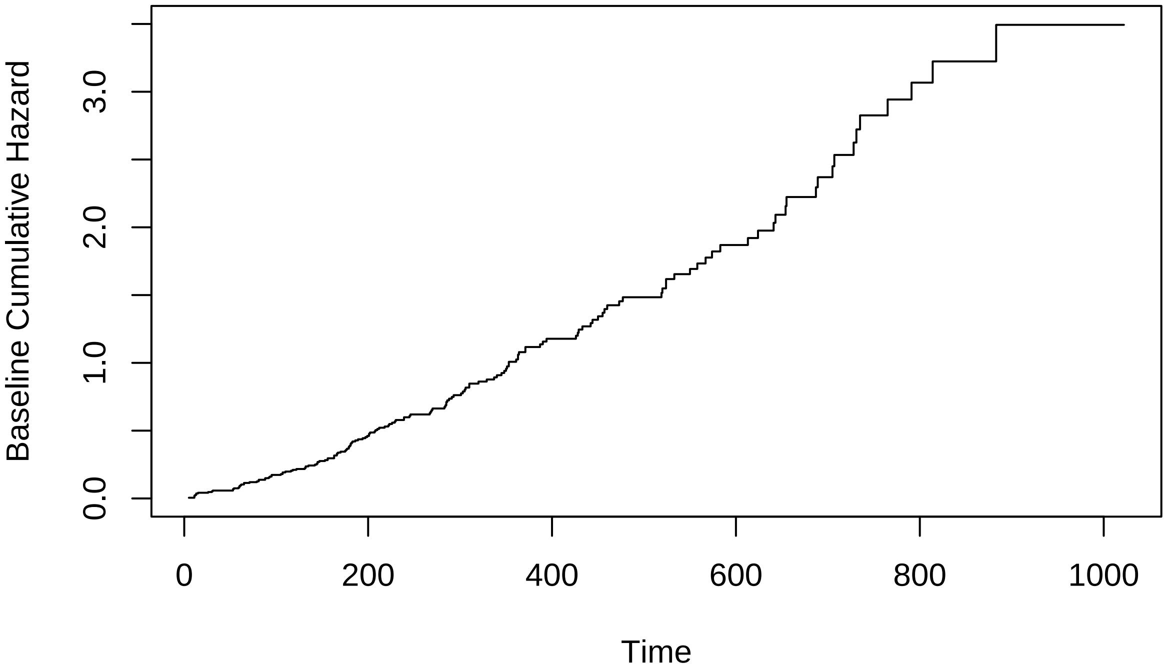
The relative risk is less than 1, implying that the hazard rate for the male aged 25 is lower than the hazard rate for the male aged 40 at entry.

The estimation of $\Lambda_0(t)$

```
out2=basehaz(out)
head(out2)
```

```
##        hazard time
## 1 0.005128541    5
## 2 0.020661842   11
## 3 0.025905907   12
## 4 0.036493728   13
## 5 0.041836471   15
## 6 0.047211324   26
```

```
plot(out2[, 2], out2[, 1], type="s",
ylab="Baseline Cumulative Hazard", xlab="Time")
```



Checking the proportionality assumption we have:

```
cox.zph(out)
```

```
##                 chisq df    p
## age             0.209  1 0.65
## as.factor(sex) 2.608  1 0.11
## GLOBAL          2.771  2 0.25
```

The proportionality assumption holds for all the covariates in the model since the p values are greater than 0.05.

We check the goodness of fit using Cox-Snell Residuals.

Cox-Snell residuals are a diagnostic tool used to assess the goodness-of-fit of a Cox proportional hazards model in survival analysis.

These residuals are particularly useful for evaluating the adequacy of the model without making specific assumptions about the baseline hazard function.

The Cox-Snell residuals are typically plotted against the observed survival times or quantiles of a theoretical distribution (e.g., exponential distribution).

A plot of Cox-Snell residuals against expected quantiles should resemble a straight line if the Cox proportional hazards model is appropriate.

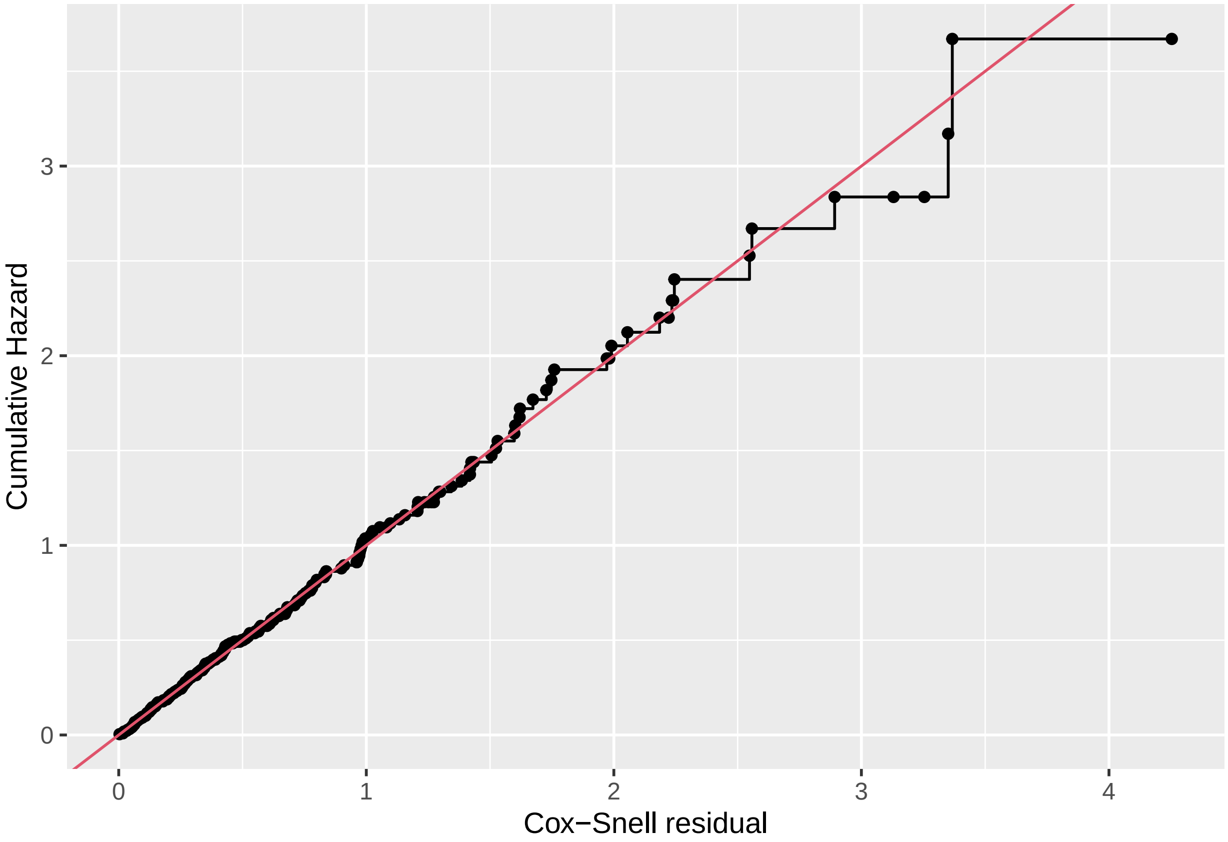
If the Cox- Snell residuals plot shows a reasonably linear pattern, it suggests that the Cox proportional hazards assumption is not violated, indicating a good fit of the model.

The linearity implies that the cumulative hazard ratio is consistent with the chosen distribution for the baseline hazard. On the other hand, departures from linearity may indicate inadequacies in the model.

```
remotes::install_github("adibender/ldatools")
```

We assess the goodness-of-fit as follows

```
library(ldatools)
library(ggplot2)
gg_coxsnell(out) +
     geom_abline(intercept=0, slope=1, col=2)
```



We observe no much deviation on the diagonal line hence the Cox-Snell. Residuals suggest that the fitted model is good.

Binomial and Poisson Models

In this lecture we introduce two more statistical models which are based on two discrete probability distributions:

- The binomial model
- The poisson model

The distributions can be used to model the number of deaths observed in a mortality investigation.

The aim of the models is to derive estimates of the true values of q_x using the binomial model or μ_x using the Poisson model.

The Binomial Model

Observe N identical, independent lives aged x exactly for one year, and record the number d who die. Then d is a sample value of a random variable D

If we suppose that each life dies with probability q_x and survives with probability $1 - q_x$, then D has a binomial distribution with parameters N and q_x

The death or survival of each life can be represented by an independent Bernoulli trial with associated probabilities of q_x and $1 - q_x$ respectively.

Here q_x refers to the *initial* rate of mortality.

For the thought out experiment outlined above, the probability that exactly d deaths will occur during the year is

$$P[D = d] = \binom{N}{d} q^d (1-q)^{N-d}$$

Proof

Since we have assumed that deaths operate independently, the probability that a specified d individuals will die during the year and the remaining $N - d$ will not is $q^d(1-q)^{N-d}$.

However we need to multiply this probability by the combinatorial factor $\binom{N}{d} = \frac{N!}{d!(N-d)!}$ which is the number of ways the d deaths would be chosen.

Estimating q_x from the data

The intuitive estimate of q_x is $\hat{q}_x = \frac{d}{N}$ and this is also the maximum likelihood estimate.

proof

Under the binomial model, the likelihood of recording exactly d deaths if the rate of mortality is q is

$$L(q) = \binom{N}{q} q^d (1-q)^{N-d}$$

which can be maximised by maximising its log

$$\log L(q) = \log \binom{N}{d} + d \log q + (N-d)\log(1-q)$$

Differentiating wrt to q

$$\frac{\partial}{\partial q} \log L(q) = \frac{d}{q} - \frac{N-d}{1-q}$$

This is zero at the value $\hat{q}$ such that:

$$d(1-\hat{q}) = (N-d)\hat{q} \Longrightarrow \hat{q} = \frac{d}{N}$$

This is maximum since

$$\frac{\partial^2}{\partial q^2}\log L(q) = -\frac{d}{q^2} - \frac{N-d}{(1-q)^2} < 0$$

The maximum likelihood estimate is the observed value of the corresponding maximum likelihood estimator

$$\tilde{q}_x = \frac{D}{N}$$

The corresponding estimator $\tilde{q}_x$ has:

- Mean q_x it is unbiased
- Variance $\frac{q_x(1-q_x)}{N}$

The observed number of deaths D has a Binomial (N, q_x) distribution which mean that it has a mean Nq_x and variance $Nq_x(1-q_x)$ so

$$E(\tilde{q}_x) = E(\frac{D}{N}) = \frac{Nq_x}{N} = q_x$$

The asymptotic distribution of $E(\tilde{q}_x)$ is

$$E(\tilde{q}_x) \sim N\Big(q_x, \frac{q_x(1-q_x)}{N}\Big)$$

This is the binomial model of mortality.

The Poisson Model

The Poisson distribution is used to model the number of ‘rare’ events occurring during some period of time.

A random variable X is said to have a Poisson distribution with mean λ if the probability function of X is

$$P(X=x) = \frac{e^{-\lambda}\lambda^x}{x!}$$

Let E_x^c denote the total observed waiting time.

If we assume that we observe N individuals as before and that the force of mortality is a constant μ, then a Poisson model is given by the assumption that D has a Poisson distribution with parameter μE_x^c That is

$$P[D=d] = \frac{e^{-\mu}E_x^c(\mu E_x^c)^d}{d!}$$

The Poisson likelihood leads to the following estimator of μ

$$\tilde{\mu} = \frac{D}{E_x^c}$$

proof

The likelihood of observing d deaths if the true value of the hazard rate is μ is

$$L(\mu) = \frac{(\mu E_x^c)^d e^{-\mu E_x^c}}{d!}$$

which can be maximised by maximising its log

$$\log L(\mu) = d(\log \mu + \log E_x^c) - \mu E_x^c - \log d!$$

Differentiating w.r.t μ

$$\frac{\partial}{\partial \mu} \log L(\mu) = \frac{d}{\mu} - E_x^c$$

which is zero when

$$\hat{\mu} = \frac{d}{E_x^c}$$

The estimator $\tilde{\mu}$ has the following properties

- $E[\tilde{\mu}] = \mu$
- $Var[\tilde{\mu}] = \frac{\mu}{E_x^c}$

The asymptotic distribution of $\tilde{\mu}$ is

$$\tilde{\mu} \sim N\Big(\mu, \frac{\mu}{E_x^c}\Big)$$

Exposed to Risk

Central exposed to risk is the total waiting time which features in both two-state markov model and the poisson model.

The central exposed to risk is a natural quantity intrinsically observable even if the observation may be incomplete in practice.

Homogeneity

The Poisson models are based on the assumption that we can observe groups of identical lives or homogeneous groups.

A group of lives with different characteristics is said to be *heterogenous*

As a result of this heterogeneity, our estimate of the mortality rate would be the estimate of the average rate over the whole group of lives.

Example

consider a country in which 50% of the population are smokers. If $\mu_{40} = 0.001$ for non-smokers and $\mu_{40} = 0.002$ for smokers, then a mortality investigation based on the entire population may lead us to the estimate $\hat{\mu}_{40} = 0.0015$ An insurance company that calculates its premiums using this average figure would overcharge non-smokers and undercharge smokers.

The solution is subdivide our data according to characteristics known, from experience, to have a significant effect on mortality. This ought to reduce the heterogeneity of each class.

Among the factors in respect of which life insurance mortality statistics are often sub-divided are:

- Sex
- Age
- Type of policy
- Smoker/non-smoker status
- Duration in force
- Level of underwriting

Principle of Correspondence

Mortality investigations based on estimation of $\mu_{x+\frac{1}{2}}$ at individual ages brings together two different items of data **deaths and exposures**

These should be defined consistently or the ratios are meaningless.

The principle of correspondence states that:

A life alive at time t should be included in the exposure at age x at time t if and only if, were that life to die immediately he or she would be counted in the death data d_x at age x.

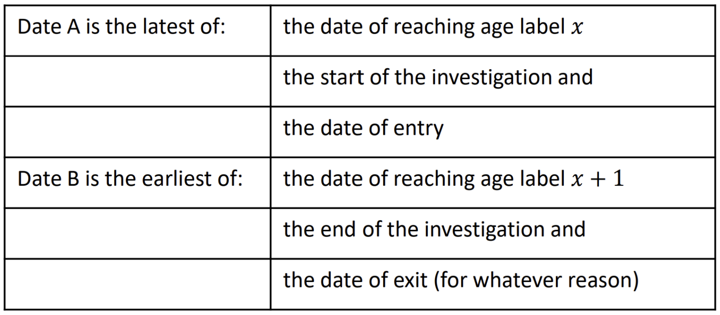
Exact Calculation of E_x^c

The procedure for the exact calculation of E_x^c is obvious:

a. Record all dates of birth
b. Record all dates of entry into observation
c. Record all dates of exit from observation
d. Compute E_x^c

If we add to the data above the cause of the cessation of observation we have d_x as well and we have finished.

The central exposed to risk E_x^c for a life with age label x is the time from Date A to Date B where

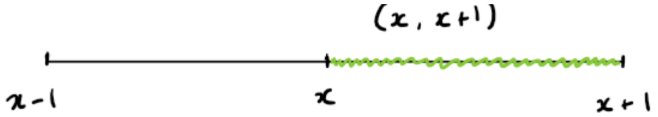


Date A is the latest of:	the date of reaching age label x
	the start of the investigation and
	the date of entry
Date B is the earliest of:	the date of reaching age label $x + 1$
	the end of the investigation and
	the date of exit (for whatever reason)

Age Definitions

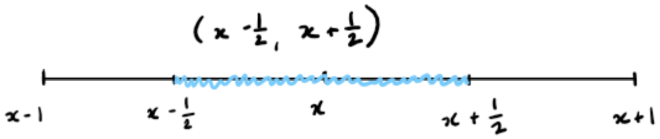
Age last birthday

A life will be considered age x with their real age being in the range $(x, x+1)$



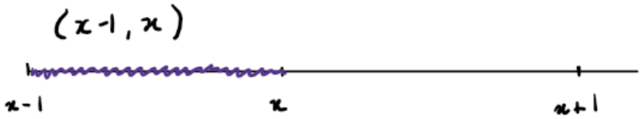
Age nearest birthday

A life will be considered age x with their real age being in the range $x - \frac{1}{2}, x + \frac{1}{2}$



Age Next Birthday

A life will be considered age x with their real age being in the range $(x-1, x)$



Census Approximation to E_x^c

Suppose that we have death data of the form:

d_x total number of deaths × last birthday during calendar years $K, K+1, ..., K+N$

That is we have over $N+1$ calendar years of all deaths between ages x and $x+1$

However, instead of the times of entry to and exit from observation of each life being known, we have instead only the following census data

$P_{x,t} $ = Number of lives under observation aged x last birthday at time t where $t = 1$ january in calendar years $K, K+1, ..., K+N, K+N+1$

Define $P_{x,t}$ to be the number of lives under observation aged x last birthday, at ant time t. Note that

$$E_x^c = \int_K^{K+N+1} P_{x,t} dt$$

During any short time interval $(t, t+dt)$ there will be $P_{x,t}$ lives each contributing a fraction of a year dt to the exposure.

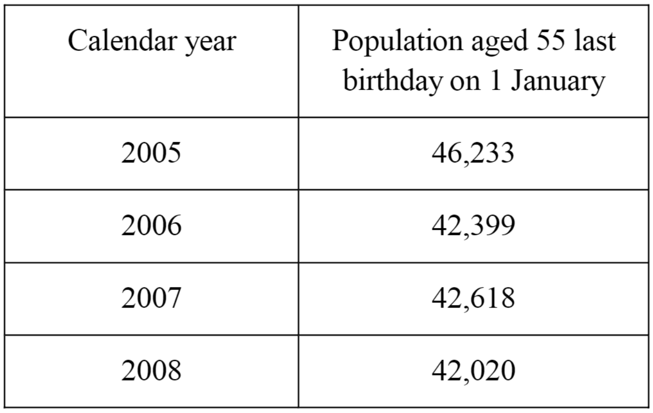
So integrating $P_{x,t} * dt$ over the observation period gives the total exposed to risk for this age.

Using the trapezium approximation

$$E_x^c = \int_K^{K+N+1} P_{x,t}dt \approx \sum_{t=K}^{K+N} \frac{1}{2}(P_{x,t} + P_{x,t+1})$$

Example

Estimate E_{55}^c based on the following data



Calendar year	Population aged 55 last birthday on 1 January
2005	46,233
2006	42,399
2007	42,618
2008	42,020

$$E_{55}^c = \int_0^3 P_{55,t}\, dt$$

$$E_{55}^c = \frac{1}{2}\left[P_{55,0} + P_{55,1}\right] + \frac{1}{2}\left[P_{55,1} + P_{55,2}\right] + \frac{1}{2}\left[P_{55,2} + P_{55,3}\right]$$

$$= \frac{1}{2}P_{55,0} + P_{55,1} + P_{55,2} + \frac{1}{2}P_{55,3}$$

$$= 0.5 * 46233 + 42399 + 42618 + 0.5 * 42020$$

$$= 129143.5$$

Deaths classified using different definitions of age

Definitions that could be used for year of age include

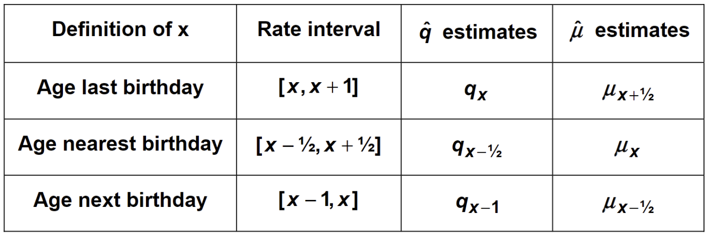
- $d_x^{(1)}$ total number of deaths at age x last birthday during calendar years $K, K+1, ..., K+N$
- $d_x^{(2)}$ total number of deaths age x nearest birthday during calendar years $K, K+1, ..., K+N$
- $d_x^{(3)}$ total number of deaths age x next birthday during calendar years $K, K+1, ..., K+N$

Rate Interval

A rate interval is a period of one year during which a life's recorded age remains the same.

The rate of mortality q measures the probability of death over the next year of age or more generally over the next rate interval.

The possibilities are:



Definition of x	Rate interval	$\hat{q}$ estimates	$\hat{\mu}$ estimates
Age last birthday	$[x, x+1]$	q_x	$\mu_{x+½}$
Age nearest birthday	$[x-½, x+½]$	$q_{x-½}$	μ_x
Age next birthday	$[x-1, x]$	q_{x-1}	$\mu_{x-½}$

Once the rate interval has been identified (from the age definition used in d_x) the rule is that

- the crude $\hat{\mu}$ estimates μ in the middle of the rate interval
- the crude $\hat{q}$ estimates q at the start of the rate interval.

Graduation and Statistical tests

Graduation refers to the process of using statistical techniques to improve the estimates provided by the crude rates.

The aims of graduation are to produce a smooth set of rates that are suitable for a particular purpose, to remove random sampling errors (as far as possible) and to use the information available from adjacent ages to improve the reliability of the estimates. Graduation results in a "smoothing" of the crude rates.

Method of Graduation

The three methods of graduation to consider include:

- Graduation by parametric formula
- Graduation by reference to a standard table
- Graphical graduation

The most appropriate method of graduation to use will depend on the quality of the data available and the purpose for which the graduated rates will be used.

Graduation by Parametric Formula

The method of graduation most often used for reasonably large experiences is to fit a parametric formula to the crude estimates.

The underlying assumption is that μ_x or q_x can be modelled using an appropriate mathematical formula with unknown parameters.

The simple but useful formulae are:

- Gompertz Formula $\mu_x = Bc^x$
- Makeham Model $\mu_x = A + Bc^x$

In practice it is usually found that μ_x follows an exponential curve quite closely over middle and older ages so most successful formulae include a Gompertz term.

Makeham's formula is interpreted as the addition of accidental deaths, independent of age to a Gompertz term representing senescent deaths.

Graduation Process

The steps include:

Select a Graduation Formula

A particular parametric family of curves must be chosen. Either:

- $$\alpha_1 \exp(\alpha_2 x)$$

  the Gompertz curve

- $$\alpha_1 + \alpha_2 \exp(\alpha_3 x)$$

  Makeham curve

- $$\alpha_1 + \alpha_2 \exp(\alpha_3 x + \alpha_4 x^2)$$

  etc

Determine Parameter Values

Use the maximum likelihood method to obtain the estimates. This is usually performed using a statistics package.

Calculate Graduated Rates

Calculate the graduated rates at each age using the fitted parametric formula. This can be done using a computer program

Test

Given the best-fitting curve of a given family, the graduated rates must be compared with the original data to see if they are acceptably close.

Graduation by Reference to Standard Table

We assume that there is a simple relationship between the observed mortality and an appropriate standard table.

Graduation by Graphical Method

We draw a curve by hand on a graph of the crude estimates.

Strengths and Weaknesses of the methods

They can be assesed through:

- Smoothness
- Precision of calculated rates
- Goodness of fit
- Ease of use
- Amount of data required
- Flexibility in allowing fo special features