# Epidemiologic Data Analysis using R
# Part 7: Analysis of survival data

Janne Pitkäniemi

Esa Läärä

Finnish Cancer Registry, Finland, <janne.pitkaniemi@cancer.fi>
University of Oulu, Finland, <esa.laara@oulu.fi>

Tampere School of Health Sciences, Jan 17-Feb 28, 2014

# Topics somewhat covered

1. Survival or time to event data & censoring.
2. Probability concepts for times to event:
   survival, hazard and cumulative hazard functions,
3. Kaplan–Meier and Nelson–Aalen estimators.
4. Regression modelling of hazards: Cox model.
5. Packages `survival`.
6. Functions `Surv()`, `survfit()`, `survMisc()`, `coxph()`.

# Survival time – time to event

Let $T$ be the **time** spent in a given **state** from its beginning till a certain *endpoint* or *outcome* **event** or *transition* occurs, changing the state to another.
(`lex.Cst - lex.dur - lex.Xst`)

Examples of such times and outcome events:

- lifetime: birth $\rightarrow$ death,
- duration of marriage: wedding $\rightarrow$ divorce,
- healthy exposure time:
  start of exposure $\rightarrow$ onset of disease,
- clinical survival time:
  diagnosis of a disease $\rightarrow$ death.

# Set-up of classical survival analysis

- **Two-state model**: only one type of event changes the initial state.
- Major applications: analysis of lifetimes since birth and of survival times since diagnosis of a disease until death from any cause.


**Censoring**: Death and final lifetime not observed for some subjects due to emigration or closing the follow-up while they are still alive

# Distribution concepts: survival function

Cumulative distribution function (CDF) $F(t)$ and density function $f(t) = F'(t)$ of survival time $T$:

$$F(t) = P(T \leq t) = \int_0^t f(u)du$$

= **risk** or probability that the event occurs by $t$.

**Survival function**

$$S(t) = 1 - F(t) = P(T > t) = \int_t^\infty f(u)du,$$

= probability of avoiding the event at least up to $t$ (the event occurs only after $t$).

# Distribution concepts: hazard function

The **hazard rate** or **intensity** function $\lambda(t)$

$$\begin{aligned}\lambda(t) &= \lim_{\Delta \to 0} P(t < T \leq t + \Delta | T > t)/\Delta \\ &= \lim_{\Delta \to 0} \frac{P(t < T \leq t + \Delta)/\Delta}{P(T > t)} = \frac{f(t)}{S(t)}\end{aligned}$$

- $\approx$ the conditional probability that the event occurs in a short interval $(t, t + \Delta]$, given that it does not occur before $t$, per interval length.

In other words, during a short interval

$$\text{risk of event} \approx \text{hazard} \times \text{interval length}$$

# Distribution: cumulative hazard etc.

The **cumulative hazard** (or integrated intensity):

$$\Lambda(t) = \int_0^t \lambda(v)dv$$

# Observed data on survival times

For individuals $i = 1, \ldots, n$ let

$T_i$ = true time to event,

$U_i$ = true time to censoring.

Censoring is assumed **noninformative**, *i.e.* independent from occurrence of events.

We observe

$y_i = \min\{T_i, U_i\}$, *i.e.* the exit time, and

$\delta_i = 1_{\{T_i < U_i\}}$, indicator (1/0) for the event occurring first, before censoring.

Censoring must properly be taken into account in the statistical analysis.

# Approaches for analysing survival time

- **Parametric models** on hazard rate $h(t)$ (like Weibull, gamma, etc.) – Likelihood:

$$L = \prod_{i=1}^{n} \lambda(y_i)^{\delta_i} S(y_i)$$

$$= \exp\left\{ \sum_{i=1}^{n} [\delta_i \ \log \ \lambda(y_i) - \Lambda(y_i)] \right\}$$

- **Piecewise constant rate** model on $\lambda(t)$ – estimation of $\hat{\lambda}(t)$ with poisson regression
- **Non-parametric** methods, like Kaplan–Meier (KM) estimator of survival curve $S(t)$ and Cox proportional hazards model.

# R package `survival`

Tools for analysis with one outcome event.

- `Surv(time,event) -> sobj`
  creates a **survival object** `sobj`,
- `survfit(sobj) -> sfo`
  non-parametric survival curve estimates, like KM (also estimated baseline in a Cox model),
- `plot(sfo)`
  survival curves and related graphs,
- `coxph(sobj ~ x1 + x2 +...)`
  fits the Cox model with covariates `x1` and `x2`.
- `survreg()` – parametric survival models.

# KM esimate for survival function $S(t)$

- Order event times (possibly separately in groups)
- $\hat{S(t_j)} = \hat{S(t_{j-1})}(1 - \frac{d_j}{n_j})$, where $t_0 = 0$ and $\hat{S(0)} = 1$
- $\hat{S(t)}$ is constant between event – step function
- Each observations contributes as long as at risk for the event and confidence intervals can be introduced using classical inference framework.

# KM esimate for survival function $S(t)$

- Order event times (possibly separately in groups)
- $\hat{S(t_j)} = \hat{S(t_{j-1})}(1 - \frac{d_j}{n_j})$, where $t_0 = 0$ and $\hat{S(0)} = 1$
- $\hat{S(t)}$ is constant between event – step function
- Each observations contributes as long as at risk for the event and confidence intervals can be introduced using classical inference framework.

# Veterans' Administration Lung Cancer study

In this trial, males with advanced inoperable lung cancer were randomized to a standard therapy and a test chemotherapy. The primary endpoint for the therapy comparison was time to death in days, represented by the variable Time.

Variables

- trt: 1=standard 2=test
- celltype:1=squamous, 2=smallcell, 3=adeno, 4=large
- time: survival time (days)
- status: status 1= death, 0=censored
- karno: Karnofsky performance score (100=good)
- diagtime:months from diagnosis to randomisation
- age: in years
- prior: prior therapy 0=no, 1=yes

Reference: D Kalbfleisch and RL Prentice (1980), The Statistical Analysis of Failure Time Data. Wiley, New York.

# Veteran data

```
> head(veteran)
  trt celltype time status karno diagtime age prior
1   1 squamous   72      1    60        7  69     0
2   1 squamous  411      1    70        5  64    10
3   1 squamous  228      1    60        3  38     0
4   1 squamous  126      1    60        9  63    10
5   1 squamous  118      1    70       11  65    10
6   1 squamous   10      1    20        5  49     0
...
70    2  squamous  999      1    90       12  54    10
71    2  squamous  112      1    80        6  60     0
72    2  squamous   87      0    80        3  48     0
73    2  squamous  231      0    50        8  52    10
74    2  squamous  242      1    50        1  70     0
```

# Estimate for hazard function $\lambda(t)$ in R – splitting time scale

```
#================================
# get packages needed for the analysis
#================================
library(dlpyr)
library(survival)
library(ggplot2)
library(Epi)
library(dummies)
#================================
# define stating time at 0 and create subject id
#================================
veteran$start<-0
veteran$id<-1:nrow(veteran)
#================================
# Split follow-up time into interval
#================================
nvet<-survSplit(
veteran,
cut=c(100,200,300,400),
event="status",
start="start",
end="time",episode="period")
```

# Estimate for hazard function $\lambda(t)$ in R – splitting time scale

```
#================================
# dplyr -package,rather than stat.table() -function
#================================
gsvet<-group_by(nvet,period)
speriod<-summarize(gsvet,
                n=length(trt),
                events=sum(status),
                pyrs=round(sum(time-start)),
                rate1000=(events/pyrs)*1000,
                lograte=log(events/pyrs))
speriod<-data.frame(speriod)
speriod<-cbind(speriod,dummy(speriod$period))
speriod
```

```
> speriod[,1:6]
  period   n events pyrs rate1000   lograte
1      0 137     79 8692 9.088817 -4.700710
2      1  53     26 3502 7.424329 -4.902993
3      2  24     10 1807 5.534034 -5.196838
4      3  13      7 1054 6.641366 -5.014438
5      4   6      6 1608 3.731343 -5.590987
```

# Estimate for hazard function $\lambda(t)$ in R – Poisson regression

```
#================================
# Summarize data with data.table, rather than stat.table() -function
#================================
gsvet<-group_by(svet,period)
speriod<-summarize(gsvet,
                n=length(trt), events=sum(status),
                pyrs=round(sum(time-start)), rate1000=(events/pyrs)*1000,
                lograte=log(events/pyrs))
speriod<-data.frame(speriod)
speriod<-cbind(speriod,dummy(speriod$period))
#================================
#  Poisson model for incidence in each period of time
#================================
m<-glm(events~1-1+speriod0+speriod1+speriod2+speriod3+speriod4,
       family=poisson,offset=log(pyrs),data=speriod)
summary(m)
round(ci.lin(m,Exp=TRUE)[,-(3:4)],6)
          Estimate   StdErr exp(Est.)     2.5%    97.5%
speriod0 -4.700710 0.112509  0.009089 0.007290 0.011331
speriod1 -4.902993 0.196116  0.007424 0.005055 0.010904
speriod2 -5.196838 0.316228  0.005534 0.002978 0.010285
speriod3 -5.014438 0.377964  0.006641 0.003166 0.013931
speriod4 -5.590987 0.408248  0.003731 0.001676 0.008306
```

# Estimation of the survival function S(t)

```
Call: survfit(formula = Surv(time, status) ~ trt, data = veteran)
                trt=1
         n         d
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    3     69       1   0.9855  0.0144      0.95771        1.000
    4     68       1   0.9710  0.0202      0.93223        1.000
    7     67       1   0.9565  0.0246      0.90959        1.000
    8     66       2   0.9275  0.0312      0.86834        0.991
   10     64       2   0.8986  0.0363      0.83006        0.973

  S(0)=1; S(3)=S(0)*(1-(1/69) )=0.985 ;S(4)=S(3)*(1-(1/68)) =0.971

                trt=2
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    1     68       2   0.9706  0.0205      0.93125        1.000
    2     66       1   0.9559  0.0249      0.90830        1.000
    7     65       2   0.9265  0.0317      0.86647        0.991
    8     63       2   0.8971  0.0369      0.82766        0.972
   13     61       1   0.8824  0.0391      0.80900        0.962

 S(0)=1; S(1)=S(0)*(1-(2/68))   =0.971; S(2)=S(1)*(1-(1/66)) =0.956
```

# Plot of Survival curve for Veteran data

```
library(survival)
library('SurvMisc')
m<-survfit(Surv(time, status)~trt, data=veteran)
summary(m)
autoplot(m,CI=TRUE,bands=TRUE,divideTime=100,plotTable=TRUE)
```

# Estimate cumulative hazard function – $\hat{\Lambda}(t)$

- KM for survival function ($P(T > t)$) is often presented
- Estimation of the cumulative hazard enables risk calculations and can be applied to competing risk situation (cancer and mortality aer competing – do not use naive KM)
- Cumulative rate - time(age) band weighted sum of hazards – estimate of $\hat{\Lambda}(t)$
- Risk: $P(T < t) = F(t) = 1 - S(t) = 1 - exp(-\Lambda(t))$
- If low incidence or short risk period:
  $1 - exp(-\Lambda(t)) \approx \Lambda(t)$ i.e. rate $\times$ period at risk
- Cumulative hazard can be estimated from KM, but Nelson-Aalen or Nelson-Johansen should be preferred

# Estimate cumulative hazard function – $\hat{\Lambda}(t)$

```
m1<-calcSurv(Surv(veteran$time[veteran$trt==1],veteran$status[veteran$trt==1]))
m2<-calcSurv(Surv(veteran$time[veteran$trt==2],veteran$status[veteran$trt==2]))
head(m1)
head(m2)
```

```
   t  n e       SKM      SKMV      HNel     SNelA     HNelV       HKM
1  3 69 1 0.9855072 0.0002070 0.0144928 0.9856118 0.0002100 0.0145988
2  4 68 1 0.9710145 0.0004079 0.0291986 0.9712235 0.0004263 0.0294139
3  7 67 1 0.9565217 0.0006027 0.0441240 0.9568353 0.0006491 0.0444518
4  8 66 2 0.9275362 0.0009741 0.0744270 0.9282752 0.0011082 0.0752234
5 10 64 2 0.8985507 0.0013211 0.1056770 0.8997152 0.0015965 0.1069721
6 11 62 1 0.8840580 0.0014855 0.1218061 0.8853200 0.0018566 0.1232326
...
t time
n no. at risk
e no. events
KM Survival estimate by Kaplan-Meier (Product-Limit) estimator
KMV  Variance of Kaplan-Meier estimate (Greenwoods formula)
SNelA Survival estimate from Nelson-Aalen estimator: S=exp(H)
HNel Nelson-Aalen estimate of cumulative hazard function
HNelV Variance of Nelson-Aalen estimate
HKM Cumulative hazard estimate from Kaplan-Meier estimator: H = -log(S)
```

# Regression models for time-to-event data

Consider only one outcome & no competing events

- Subject $i$ ($i = 1, \ldots, n$) has an own vector $x_i$ that contains values $(x_{i1}, \ldots, x_{ip})$ of a set of $p$ continuous and/or binary covariate terms.
- In the spirit of generalized linear models we let $\beta = (\beta_1, \ldots, \beta_p)$ be regression coefficients and build a **linear predictor**

$$\eta_i = x_i^{\mathsf{T}} \beta = \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$$

- Specification of outcome variable? Distribution (family)? Expectation? Link?

# Regression models (cont'd)

Survival regression models can be defined *e.g.* for

(a) survival times directly

$$\log(T_i) = \eta_i + \epsilon_i, \quad \text{s.t. } \epsilon_i \sim F_0(t; \alpha)$$

where $F_0(t; \alpha)$ is some baseline model,

(b) hazards, multiplicatively:

$$\lambda_i(t) = \lambda_0(t; \alpha) r(\eta_i), \quad \text{where}$$

$\lambda_0(t; \alpha) =$ baseline hazard and
$r(\eta_i) =$ relative rate function, typically $\exp(\eta_i)$

(c) hazards, additively:

$$\lambda_i(t) = \lambda_0(t; \alpha) + \eta_i.$$

# Relative hazards model or Cox model

In model (b), the baseline hazard $\lambda_0(t, \alpha)$ may be given a parametric form (*e.g.* Weibull) or a piecewise constant rate (exponential) structure.

Often a parameter-free form $\lambda_0(t)$ is assumed. Then

$$\lambda_i(t) = \lambda_0(t) \exp(\eta_1),$$

specifies the **Cox model** or the **semiparametric proportional hazards model**.

$\eta_i = \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$ not depending on time.

Generalizations: **time-dependent** covariates $x_{ij}(t)$, and/or effects $\beta_j(t)$.

# PH model: interpretation of parameters

Present the model explicitly in terms of $x$'s and $\beta$'s.

$$\lambda_i(t) = \lambda_0(t) \exp(\beta_1 x_{i1} + \cdots + \beta_p x_{ip})$$

Consider two individuals, $i$ and $i'$, having the same values of all other covariates except the $j^{\text{th}}$ one.

The ratio of hazards is constant:

$$\frac{\lambda_i(t)}{\lambda_{i'}(t)} = \frac{\exp(\eta_i)}{\exp(\eta_{i'})} = \exp\{\beta_j(x_{ij} - x_{i'j})\}.$$

Thus $e^{\beta_j} = \text{HR}_j =$ **hazard ratio** or relative rate associated with a unit change in covariate $X_j$.

# Ex. Veteran data and treatment effect

Fitting Cox models with trt effect.

```
Call:
coxph(formula = Surv(time, status) ~ trt, data = veteran)

  n= 137, number of events= 128

       coef exp(coef) se(coef)     z Pr(>|z|)
trt 0.01774   1.01790  0.18066 0.098    0.922

    exp(coef) exp(-coef) lower .95 upper .95
trt     1.018     0.9824    0.7144      1.45

Concordance= 0.525  (se = 0.026 )
Rsquare= 0   (max possible= 0.999 )
Likelihood ratio test= 0.01  on 1 df,   p=0.9218
Wald test            = 0.01  on 1 df,   p=0.9218
Score (logrank) test = 0.01  on 1 df,   p=0.9218
```

HR for treatment 2 vs. 1 is 1.02 (95% CI 0.71;1.45)

Not statistically significant (p=0.92)

# Proportionalilty of hazards?

- Consider two groups $g$ and $h$ defined by one categorical covariate, and let $\rho > 0$.

  If $\lambda_g(t) = \rho\lambda_h(t)$ then $\Lambda_g(t) = \rho\Lambda_h(t)$ and

  $$\log\ \Lambda_g(t) = \log(\rho) + \log\ \Lambda_h(t),$$

  thus log-cumulative hazards should be parallel!

$\Rightarrow$ *Plot the estimated log-cumulative hazards and see whether they are sufficiently parallel*.

- `plot(coxobj, ..., fun = 'cloglog')`
- Testing the proportionality assumptions: `cox.zph(coxobj)`.

# Ex. Veteran data - test PH

- With $> 1$ covariates, `cox.zph()` tests the assumption by checking, whether the corresponding parameters (& hazard ratios) may vary in time.
- Suppose that I want to include information on patient baseline general disease status – Karnofsky performance score (0=dead, 100=good)
- Dichotomize the Karnofsky score 0 if Score [0,50 ] and 1 if (50,100]

# Ex. Veteran data - test PH

```
> veteran$karnod<-as.numeric(veteran$karno>50)
> m<-coxph(formula = Surv(time, status) ~ trt+ karnod, data = veteran)
> summary(m)
  n= 137, number of events= 128

           coef exp(coef) se(coef)      z Pr(>|z|)
trt     0.1176    1.1248   0.1826  0.644    0.519
karnod -0.9790    0.3757   0.1895 -5.165  2.4e-07 ***

       exp(coef) exp(-coef) lower .95 upper .95
trt       1.1248      0.889    0.7864    1.6088
karnod    0.3757      2.662    0.2591    0.5447

#testing proportionality
> cox.zph(m)
           rho chisq        p
trt    -0.152   3.1 7.85e-02
karnod  0.365  15.2 9.82e-05
GLOBAL     NA  16.7 2.33e-04
```

Test for proportionality are significant ($p<0.05$) – assumptions of proportionalty of hazards is rejected for both tretment and score variable – stratify accoding to the score

# Ex. Veteran data - test PH

```
> m<-coxph(formula = Surv(time, status) ~ trt+strata(karnod),
+          data = veteran)
> summary(m)
Call:
coxph(formula = Surv(time, status) ~ trt + strata(karnod), data = veteran)

  n= 137, number of events= 128

       coef exp(coef) se(coef)     z Pr(>|z|)
trt 0.01351   1.01360  0.18372 0.074    0.941

    exp(coef) exp(-coef) lower .95 upper .95
trt     1.014     0.9866    0.7071     1.453

> cox.zph(m)
       rho chisq     p
trt -0.133  2.31 0.129
```

Test for proportionality is not significant (p>0.05)
HR for treatment effect is 1.01 95% CI (0.71 ;1.45)

# Homogeneity of HRs

Question: Are the HRs for different celltypes similar or not ?

Testing hypothesis of regression coefficients equal (Not the hypothesis that they are zero)

More formally, the model:

$$\lambda_i(t) = \lambda_0(t) \exp(\beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3})$$

, where $x_{i1} = 1$ if celltype smallcell 0 otherwise
and $x_{i2} = 1$ if celltype adeno 0 otherwise
and $x_{i3} = 1$ if celltype large 0 otherwise
and squamous celltype has been chosen as the reference category ($x_{i1} = 0$ and $x_{i2} = 0$ and $x_{i3} = 0$)

# Homogeneity of HRs

Testing hypothesis of regression coefficients equal (Not the hypothesis that they are zero)

Hypothesis more formally:

$$H_0 : \beta_1 = \beta_2 = \beta_3$$

$$H_A : \beta_1 \neq \beta_2 \text{ and } \beta_2 \neq \beta_3$$

Note that this test is different from testing hypothesis

$$H_0 : \beta_1 = 0 \text{ and } \beta_2 = 0 \text{ and } \beta_3 = 0$$

# Homogeneity of HRs

```
> m<-coxph(formula = Surv(time, status) ~ celltype,
+          data = veteran)
> summary(m)
Call:
coxph(formula = Surv(time, status) ~ celltype, data = veteran)

  n= 137, number of events= 128

                    coef exp(coef) se(coef)     z Pr(>|z|)
celltypesmallcell 1.0013    2.7217   0.2535 3.950 7.83e-05 ***
celltypeadeno     1.1477    3.1510   0.2929 3.919 8.90e-05 ***
celltypelarge     0.2301    1.2588   0.2773 0.830    0.407
---

                  exp(coef) exp(-coef) lower .95 upper .95
celltypesmallcell     2.722     0.3674     1.656     4.473
celltypeadeno         3.151     0.3174     1.775     5.594
celltypelarge         1.259     0.7944     0.731     2.168

Concordance= 0.608  (se = 0.029 )
Rsquare= 0.166   (max possible= 0.999 )
Likelihood ratio test= 24.85  on 3 df,   p=1.661e-05
Wald test            = 24.09  on 3 df,   p=2.387e-05
Score (logrank) test = 25.51  on 3 df,   p=1.208e-05
```

# Homogeneity of HRs

Let's look first are the HRs for celltypes pairwisely similar

```
> linearHypothesis(m,c("celltypesmallcell=celltypeadeno"))
Linear hypothesis test
Hypothesis: celltypesmallcell - celltypeadeno = 0

Model 1: restricted model
Model 2: Surv(time, status) ~ celltype

  Res.Df Df  Chisq Pr(>Chisq)
1    135
2    134  1 0.3452     0.5569
> linearHypothesis(m,c("celltypeadeno=celltypelarge"))
Linear hypothesis test
Hypothesis: celltypeadeno - celltypelarge = 0

Model 1: restricted model
Model 2: Surv(time, status) ~ celltype

  Res.Df Df  Chisq Pr(>Chisq)
1    135
2    134  1 10.153   0.001441 **
```

The HRs for Adeno and large celltype cancers are not significantly (p=0.56), but adeno and large type are when

# Homogeneity of HRs

Joint the for homogeneity of HRs

```
> linearHypothesis(m,c("celltypesmallcell=celltypeadeno",
+                      "celltypeadeno=celltypelarge")
+                    )
Linear hypothesis test

Hypothesis:
celltypesmallcell - celltypeadeno = 0
celltypeadeno - celltypelarge = 0

  Res.Df Df  Chisq Pr(>Chisq)
1    136
2    134  2 12.328   0.002104 **
```

HR's are not the same between celltypes

# Predictions from the Cox model

- Individual survival *times* cannot be predicted but ind'l survival *curves* can. PH model implies:

$$S_i(t) = [S_0(t)]^{\exp(\beta_1 x_{i1} + \ldots + \beta_p x_{ip})}$$

- Having estimated $\beta$ by partial likelihood, the baseline $S_0(t)$ is estimated by Breslow method
- From these, a survival curve for an individual with given covariate values is predicted.
- In R: `pred <- survfit(mod, newdata=...)` and `plot(pred)`, where `mod` is the fitted `coxph` object, and `newdata` specifies the covariate values.

# Modelling with competing risks

Main options, providing answers to different questions.

(a) Cox model for event-specific hazards
$h_c(t) = f_c(t)/[1 - F(t)]$, when *e.g.* the interest is in the biological effect of the prognostic factors on the fatality of the very disease that often leads to the relevant outcome.

(b) **Fine–Gray model** for the hazard of the subdistribution
$\gamma_c(t) = f_c(t)/[1 - F_c(t)]$ when we want to assess the impact of the factors on the overall cumulative incidence of event $c$.
– Function `crr()` in package `cmprsk`.