Statistical Methods in Cancer Epidemiology using R

Karri Seppä

Finnish Cancer Registry

Lecture 7

karri.seppa@cancer.fi

Mar 9, 2020

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.

Main R functions to be covered (`survival` package)

- `Surv()`, `survfit()`, `survminer()`, `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 $\lambda(t)$

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

- $\approx$ 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

$$T_i = \text{true time to event,}$$

$$U_i = \text{true time to censoring.}$$

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)
- $\widehat{S(t_j)} = \widehat{S(t_{j-1})}\left(1 - \frac{d_j}{n_j}\right)$, where $t_0 = 0$ and $\widehat{S(0)} = 1$
- $\widehat{S(t)}$ 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.

Age-standardised esimate for survival function $S(t)$

- Weighted average of age-specific survival estimates

$$S(t) = \sum_{a=1}^{K} w_a S_a(t) \quad \text{where } \sum_{a=1}^{K} w_a = 1$$

- weight w_a is a standard for the proportion of patients in age group a at the beginning of follow-up
  - e.g. the international cancer survival standards (ICSS; Corazziari et al. 2004)
- Can be estimated using `survtab()` function in `popEpi` package

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

```
library(survival)
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
```

```
head(veteran[veteran$trt==2,])
```

```
   trt celltype time status karno diagtime age prior
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
75   2 squamous  991      1    70        7  50    10
```

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

```
#================================
# get packages needed for the analysis
#================================
library(dplyr)
library(survival)
library(Epi)
#================================
# 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[,1:6]
```

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

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

```
#================================
# Poisson model for incidence in each period of time
#================================
m<-glm(events ~ -1+factor(period),
       family=poisson,offset=log(pyrs),data=speriod)
round(ci.lin(m,Exp=TRUE)[,-(3:4)],6)
```

```
                  Estimate   StdErr exp(Est.)     2.5%    97.5%
factor(period)1 -4.700710 0.112509   0.009089 0.007290 0.011331
factor(period)2 -4.902993 0.196116   0.007424 0.005055 0.010904
factor(period)3 -5.196838 0.316228   0.005534 0.002978 0.010285
factor(period)4 -5.014438 0.377964   0.006641 0.003166 0.013931
factor(period)5 -5.590987 0.408248   0.003731 0.001676 0.008306
```

```
#1-year cumulative hazard
Lambda_1yr<- sum(exp(m$coefficient[1:4])*c(100,100,100,65))
#1-year survival in percentages
exp(-Lambda_1yr)*100
```

```
[1] 7.161814
```

Estimation of the survival function $S(t)$

```
m <- survfit(formula = Surv(time, status) ~ trt, data = veteran)
m2 <- summary(m)
m2 <- data.frame(lapply(c(10,2:6) , function(x) m2[x]))
m2[m2$time<10,]
```

```
   strata time n.risk n.event n.censor      surv
1   trt=1    3     69       1        0 0.9855072
2   trt=1    4     68       1        0 0.9710145
3   trt=1    7     67       1        0 0.9565217
4   trt=1    8     66       2        0 0.9275362
58  trt=2    1     68       2        0 0.9705882
59  trt=2    2     66       1        0 0.9558824
60  trt=2    7     65       2        0 0.9264706
61  trt=2    8     63       2        0 0.8970588
```

- trt=1

  $S(0) = 1$

  $S(3) = S(0) \times (1 - 1/69) = 0.986$

  $S(4) = S(3) \times (1 - 1/68) = 0.971$

- trt=2

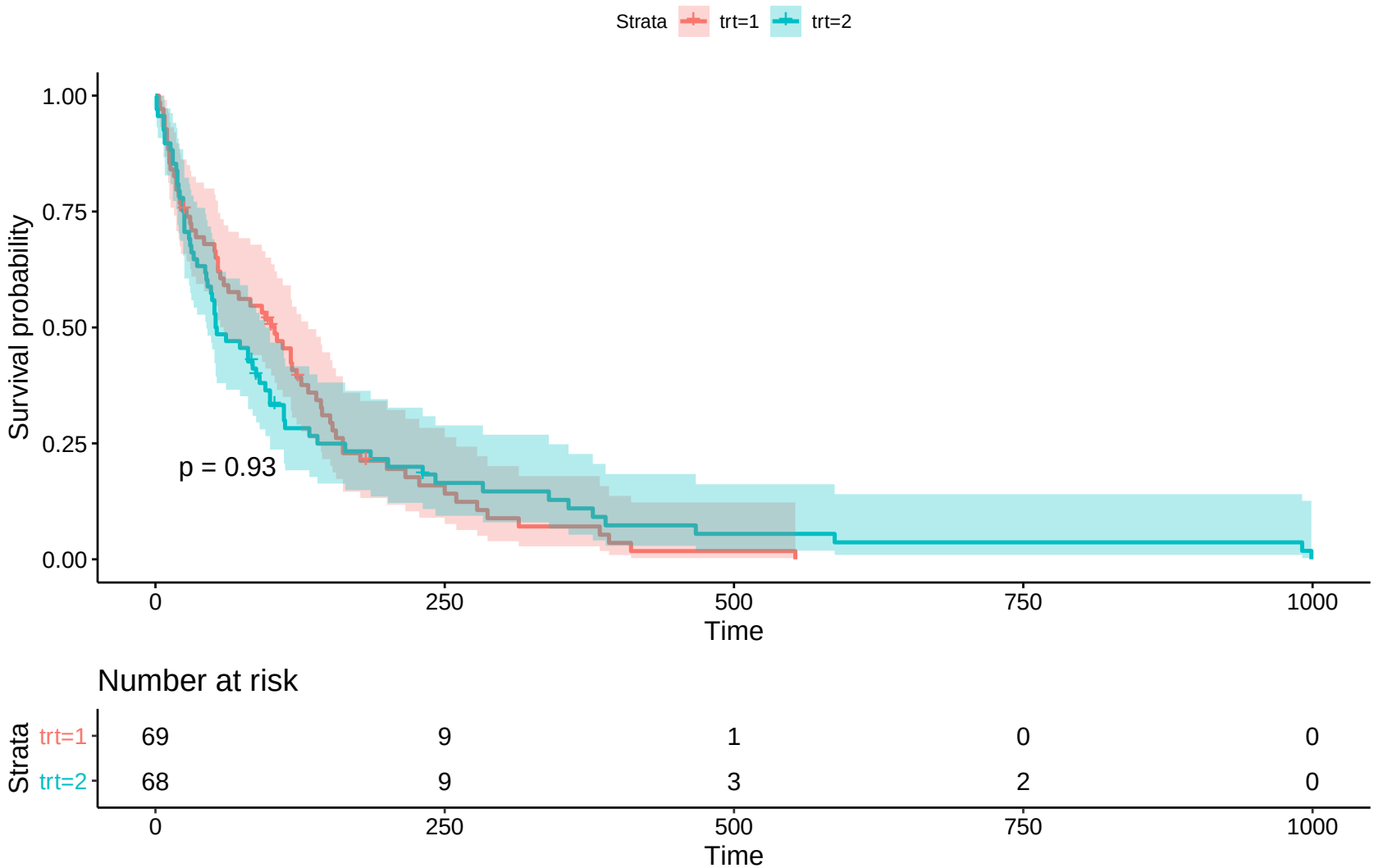
  $S(0) = 1$

  $S(1) = S(0) \times (1 - 2/68) = 0.971$

  $S(2) = S(1) \times (1 - 1/66) = 0.956$

Plot of Survival curve for Veteran data

```
library(survminer)
ggsurvplot(m, pval = TRUE, conf.int = TRUE, risk.table = TRUE,
           risk.table.y.text.col = TRUE)
```

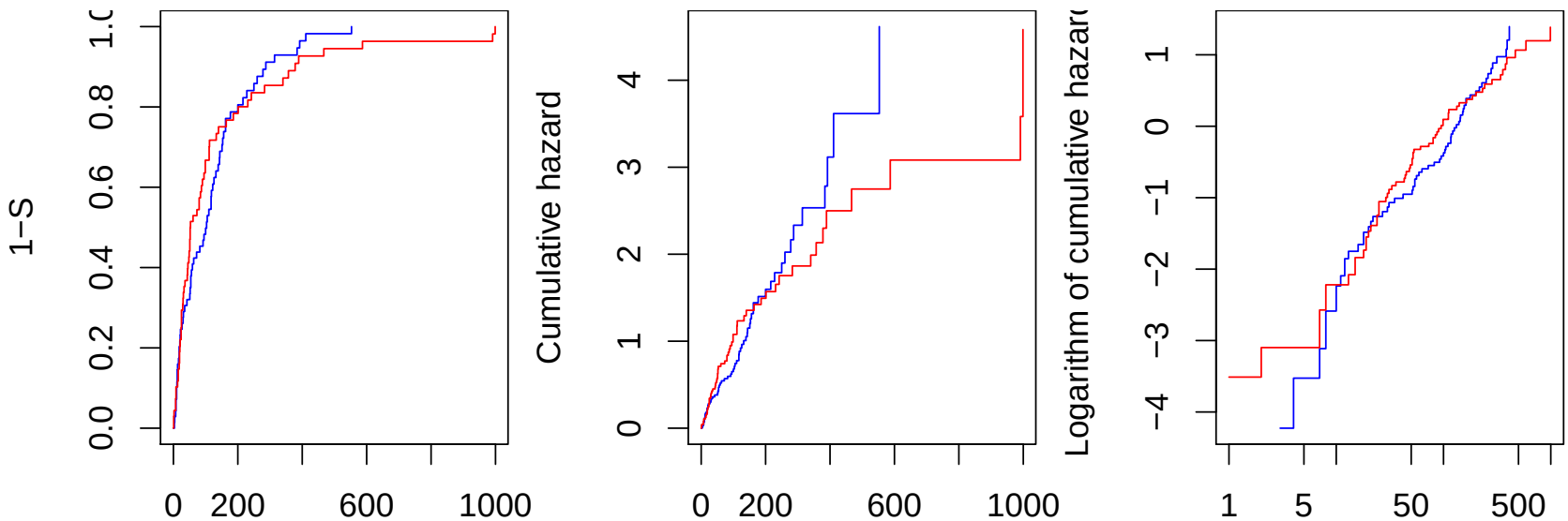


Estimate cumulative hazard function – $\hat{\Lambda}(t)$

- KM for survival function $P(T > t)$ is often presented
  $P(T > t) = S(t) = \exp[-\Lambda(t)] = \exp\left[-\int_0^t \lambda(u)du\right]$
- Cumulative risk from 0 to t:
  $P(T \leq t) = 1 - S(t) = 1 - \exp[-\Lambda(t)]$
- If low incidence rate λ or short risk period Δ:
  $1 - \exp[-\Lambda(t)] \approx \Lambda(t) = \lambda\Delta$ i.e. rate $\times$ period at risk
- Cumulative hazard can be estimated from KM, but Nelson-Aalen should be preferred

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

```
m <- survfit(formula = Surv(time, status) ~ trt, data = veteran)
par(mfrow=c(1,3), mar=c(3,4,0,0.5))
plot(m, fun="F",ylab="1-S", col=c("blue","red"))
plot(m, fun="cumhaz", ylab="Cumulative hazard", col=c("blue","red"))
plot(m, fun="cloglog", ylab="Logarithm of cumulative hazard", col=c("blue","red
```



- KM curve of survival S(t) is the most popular.
- Informative are also graphs for estimates of
  - $F(t) = 1 - S(t)$, i.e. CDF
  - $\Lambda(t) = -\log[1 - F(t)]$, cumulative hazard,
  - $\log[\Lambda(t)]$, cloglog transform of CDF.

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 β'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^{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.

```
m <- coxph(formula = Surv(time, status) ~ trt, data = veteran)
summary(m)
```

```
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 )
Likelihood ratio test= 0.01  on 1 df,   p=0.9
Wald test            = 0.01  on 1 df,   p=0.9
Score (logrank) test = 0.01  on 1 df,   p=0.9
```

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)
m
```

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

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

Likelihood ratio test=24.66  on 2 df, p=4.426e-06
n= 137, number of events= 128
```

```
#testing proportionality
cox.zph(m)
```

```
       chisq df       p
trt     1.69  1    0.19
karnod 16.04  1 6.2e-05
GLOBAL 20.21  2 4.1e-05
```

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)
m
```

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

       coef exp(coef) se(coef)     z     p
trt 0.01351   1.01360  0.18372 0.074 0.941

Likelihood ratio test=0.01  on 1 df, p=0.9414
n= 137, number of events= 128
```

```
cox.zph(m)
```

```
       chisq df    p
trt     2.56  1 0.11
GLOBAL  2.56  1 0.11
```

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

```
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
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                  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 = 0.028 )
Likelihood ratio test= 24.85  on 3 df,   p=2e-05
Wald test            = 24.09  on 3 df,   p=2e-05
Score (logrank) test = 25.51  on 3 df,   p=1e-05
```

Homogeneity of HRs

Test for homogeneity across HRs of three cell types: small cell, adeno and large ($\beta_1 = \beta_2 = \beta_3$?)

```
veteran$celltype2 <- Relevel(veteran$celltype,
                             list(1,"small_or_adeno_or_large"=2:4))
m2<-coxph(formula = Surv(time, status) ~ celltype2,
          data = veteran)
anova(m,m2)
```

```
Analysis of Deviance Table
 Cox model: response is  Surv(time, status)
 Model 1: ~ celltype
 Model 2: ~ celltype2
   loglik  Chisq Df P(>|Chi|)
1 -493.02
2 -499.76 13.475  2  0.001186 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

HR's are not the same between celltypes

Homogeneity of HRs

Test for homogeneity between HRs of smallcell and adeno cancer ($\beta_2 = \beta_3$?).

```
veteran$celltype3 <- Relevel(veteran$celltype,
                             list(1,"small_or_adeno"=2:3,4))
m3<-coxph(formula = Surv(time, status) ~ celltype3,
          data = veteran)
anova(m,m3)
```

```
Analysis of Deviance Table
 Cox model: response is  Surv(time, status)
 Model 1: ~ celltype
 Model 2: ~ celltype3
   loglik  Chisq Df P(>|Chi|)
1 -493.02
2 -493.20 0.3407  1     0.5594
```

The HRs do not differ significantly between small cell and adeno cancers (p=0.56).