Analiza danych

prof. UAM dr hab. Tomasz Górecki

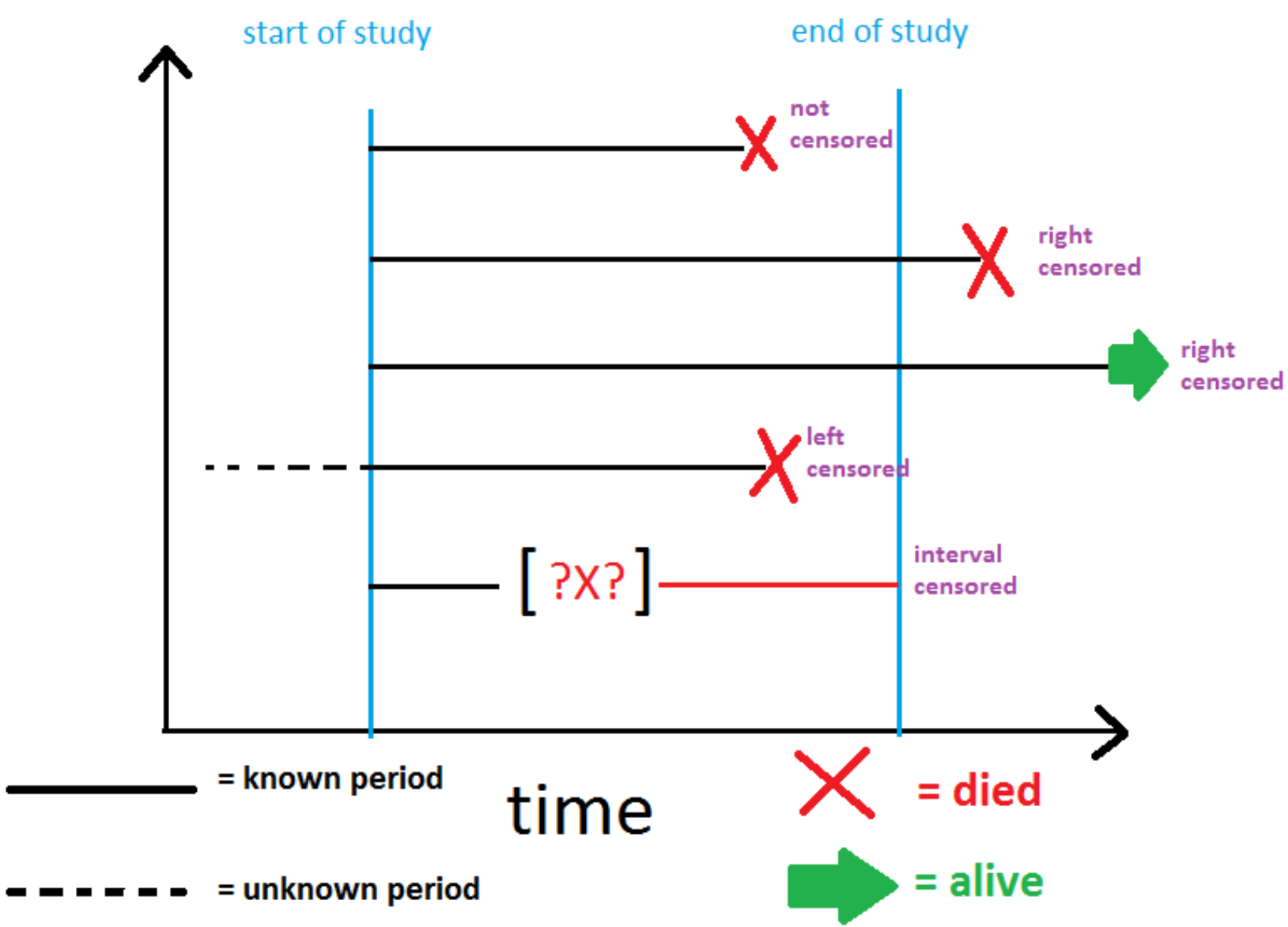
tomasz.gorecki@amu.edu.pl
http://drizzt.home.amu.edu.pl

Zakład Statystyki Matematycznej i Analizy Danych
Wydział Matematyki i Informatyki
Uniwersytet im. Adama Mickiewicza w Poznaniu

Wprowadzenie

Przedmiotem badania analizy przeżycia (*ang. survival analysis*) jest czas jaki upływa od początku obserwacji do wystąpienia określonego zdarzenia, które jednoznacznie kończy obserwację na danej jednostce eksperymentalnej. Analiza przeżycia wywodzi się wprawdzie z badań medycznych (na co wskazuje nazwa), lecz znajduje również zastosowanie w innych naukach (np. analiza niezawodności w naukach technicznych). Charakterystyczne dla analizy przeżycia są tzw. dane cenzorowane (*ang. censored observations*) inaczej ucięte (oznaczamy 70^+), o których wiadomo, że zdarzenie nie nastąpiło aż do momentu zakończenia obserwacji (np. pacjenci wypisani ze szpitala).

Wprowadzenie



Wprowadzenie

Głównym obiektem badawczym jest tzw. funkcja przeżycia (*ang. survival function*) $S(t)$, która określa prawdopodobieństwo, że osoba przeżyje dłużej niż pewien przyjęty czas t, czyli

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

gdzie T jest absolutnie ciągłą zmienną losową określającą czas życia o funkcji gęstości f oraz dystrybuancie F i przyjmującą wartości z przedziału $[0, \infty)$

Estymator Kaplana-Meiera

Najpopularniejszym estymatorem funkcji przeżycia jest estymator KAPLANA-MEIERA (*ang. Kaplan's Meier estimator*) postaci:

$$\hat{S}(t) = \prod_{t_i \leq t} \frac{n_i - d_i}{n_i},$$

gdzie n_i jest liczbą obiektów, która dożyła momentu t_i, natomiast d_i jest liczbą śmierci w momencie t_i. Ważną jego zaletą jest uwzględnianie obserwacji cenzorowanych.

Estymator Kaplana-Meiera – przykład

Znane są czasy przeżycia (w dniach) dla 16 pacjentów, którzy przeszli zabieg usunięcia guza mózgu: 28, 49, 54, 80, 80, 102^+, 120, 120^+, 120^+, 167, 200, 200, 200^+, 340, 500, 500^+.
Znajdziemy dla tych danych estymator KM funkcji przeżycia.
Załóżmy wpierw, że nie uwzględniamy cenzorowania danych.

i	t_i	n_i	d_i	$\frac{n_i - d_i}{n_i}$	$\hat{S}(t = t_i)$
1	28	16	1	0.9375	0.9375
2	49	15	1	0.9333	0.8750
3	54	14	1	0.9286	0.8125
4	80	13	2	0.8462	0.6875
5	102	11	1	0.9091	0.6250
6	120	10	3	0.7000	0.4375
7	167	7	1	0.8571	0.3750
8	200	6	3	0.5000	0.1875
9	340	3	1	0.6667	0.1250
10	500	2	2	0.0000	0.0000

Estymator Kaplana-Meiera – przykład

Z kolei jeśli uwzględnimy, że cześć obserwacji była cenzorowana:

i	t_i	n_i	d_i	$\frac{n_i - d_i}{n_i}$	$\hat{S}(t = t_i)$
1	28	16	1	0.9375	0.9375
2	49	15	1	0.9333	0.8750
3	54	14	1	0.9286	0.8125
4	80	13	2	0.8462	0.6875
5	120	10	1	0.9000	0.6188
6	167	7	1	0.8571	0.5304
7	200	6	2	0.6667	0.3536
8	340	3	1	0.6667	0.2357
9	500	2	1	0.5000	0.1179

Dane: 28, 49, 54, 80, 80, 102^+, 120, 120^+, 120^+, 167, 200, 200, 200^+, 340, 500, 500^+.

Model Coxa

Estymator KM pozwala na graficzną prezentację krzywej przeżycia oraz porównanie takich krzywych dla kilku populacji. Nie jest jednak możliwe za jego pomocą opisanie zależności funkcji przeżycia od innych zmiennych objaśniających. Wydawać by się mogło, że do tego zagadnienia powinniśmy zastosować poznane już modele regresji wielokrotnej. Nie jest to jednak możliwe z dwóch powodów: czas przeżycia przeważnie nie ma rozkładu normalnego oraz występuje problem wykorzystania obserwacji cenzorowanych.

Model Coxa

Najpopularniejszym modelem regresji wykorzystywanym w tym przypadku jest model proporcjonalnego hazardu Coxa (*ang. Cox proportional hazard model*). Zdefiniujmy wpierw funkcję hazardu (*ang. hazard function*):

$$\lambda(t) = \lim_{dt \to 0} \frac{\mathbb{P}(t \leq T \leq t + dt)}{dt \cdot S(t)} = \frac{f(t)}{S(t)} = \frac{S'(t)}{S(t)}.$$

Funkcja hazardu nie jest prawdopodobieństwem, ale miarą ryzyka śmierci w chwili t (im większy hazard tym większe ryzyko).

Każda nieujemna funkcja λ, dla której zachodzi

$$\int_0^\infty \lambda(t)dt = \infty$$

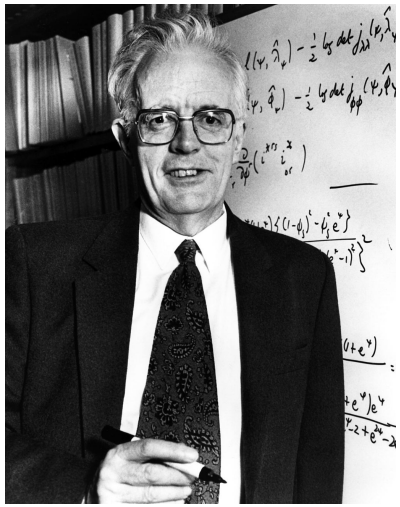
może być funkcją hazardu.

Model Coxa

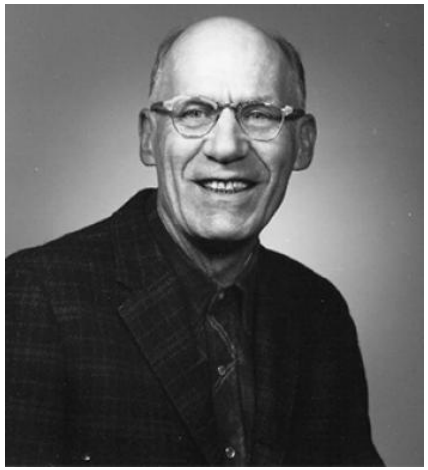
Model proporcjonalnego hazardu Coxa ma postać:

$$\lambda(t \mid \boldsymbol{X}) = \lambda_0(t) e^{\boldsymbol{X}\boldsymbol{\beta}},$$

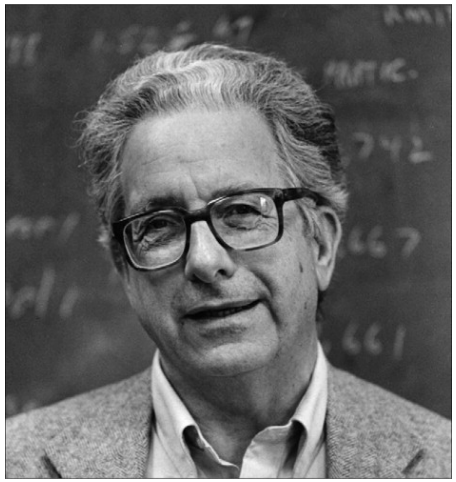
gdzie $\lambda_0(t)$ jest hazardem zerowym (*ang. baseline hazard function*) – hazard, gdy wszystkie zmienne niezależne są równe zero. W modelu tym nie zakładamy nic o postaci funkcji hazardu (takie założenie można wprowadzić uzyskując parametryczne modele proporcjonalnego hazardu). Jeśli założymy, że elementy wektora parametrów nie zależą od czasu, to mamy do czynienia z modelem proporcjonalnego hazardu Coxa.



Sir David Roxbee Cox
(1924-2022)



Edward Lynn Kaplan
(1920-2006)



Paul Meier
(1924-2011)

Bibliografia

Cox, D.R. (1972). Regression models and life tables. *Journal of the Royal Statistical Society: Series B* 34(2):187–220.

Kaplan, E.L., Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* 53(282):457–481.

Biblioteka survminer – karta pomocy

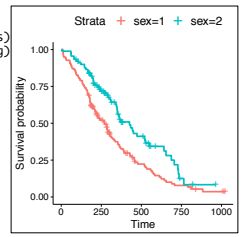
Creating Survival Plots

Informative and Elegant with *survminer*

Survival Curves

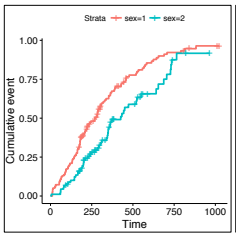
The **ggsurvplot()** function creates **ggplot2** plots from **survfit** objects.

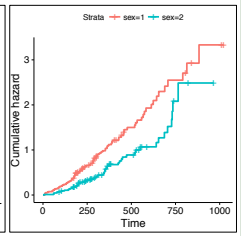
```
library("survival")
fit <- survfit(Surv(time,status)
           ~ sex, data = lung)

class(fit)
## [1] "survfit"

library("survminer")
ggsurvplot(fit, data = lung)
```



Use the **fun** argument to set the transformation of the survival curve. E.g. **"event"** for cumulative events, **"cumhaz"** for the cumulative hazard function or **"pct"** for survival probability in percentage.

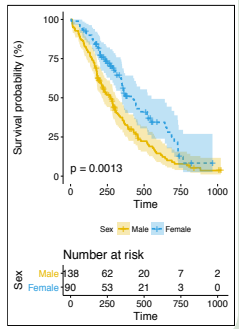
```
ggsurvplot(fit, data = lung, fun = "event")
ggsurvplot(fit, data = lung, fun = "cumhaz")
```





With lots of graphical parameters you have full control over look and feel of the survival plots; position and content of the legend; additional annotations like p-value, title, subtitle.

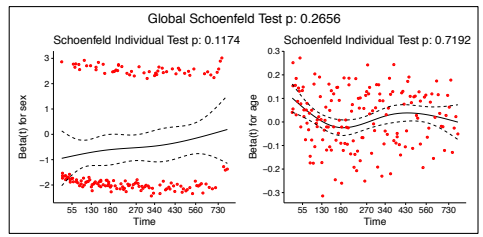
```
ggsurvplot(fit, data = lung,
      conf.int = TRUE,
      pval = TRUE,
      fun = "pct",
      risk.table = TRUE,
      size = 1,
      linetype = "strata",
      palette = c("#E7B800",
                  "#2E9FDF"),
      legend = "bottom",
      legend.title = "Sex",
      legend.labs = c("Male",
                      "Female"))
```



Diagnostics of Cox Model

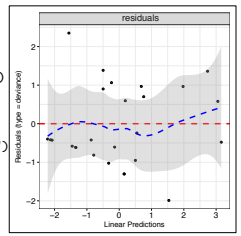
The function **cox.zph()** from **survival** package may be used to test the proportional hazards assumption for a Cox regression model fit. The graphical verification of this assumption may be performed with the function **ggcoxzph()** from the **survminer** package. For each covariate it produces plots with scaled Schoenfeld residuals against the time.

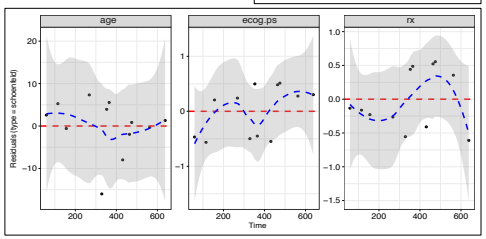
```
library("survival")
fit <- coxph(Surv(time, status) ~ sex + age, data = lung)
ftest <- cox.zph(fit)
ftest
##          rho chisq     p
## sex    0.1236 2.452 0.117
## age   -0.0275 0.129 0.719
## GLOBAL     NA 2.651 0.266
library("survminer")
ggcoxzph(ftest)
```



The function **ggcoxdiagnostics()** plots different types of residuals as a function of time, linear predictor or observation id. The type of residual is selected with **type** argument. Possible values are "martingale", "deviance", "score", "schoenfeld", "dfbeta"', "dfbetas", and "scaledsch".
The **ox.scale** argument defines what shall be plotted on the OX axis. Possible values are "linear.predictions", "observation.id", "time".
Logical arguments **hline** and **sline** may be used to add horizontal line or smooth line to the plot.

```
library("survival")
library("survminer")
fit <- coxph(Surv(time, status) ~
          sex + age, data = lung)

ggcoxdiagnostics(fit,
      type = "deviance",
      ox.scale = "linear.predictions")

ggcoxdiagnostics(fit,
      type = "schoenfeld",
      ox.scale = "time")
```

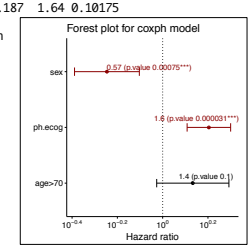




Summary of Cox Model

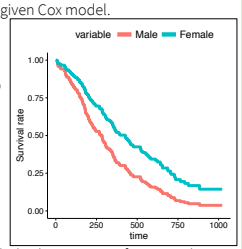
The function **ggforest()** from the **survminer** package creates a forest plot for a Cox regression model fit. Hazard ratio estimates along with confidence intervals and p-values are plotter for each variable.

```
library("survival")
library("survminer")
lung$age <- ifelse(lung$age > 70, ">70","<= 70")
fit <- coxph( Surv(time, status) ~ sex + ph.ecog + age, data = lung)
fit

## Call:
## coxph(formula = Surv(time, status) ~ sex+ph.ecog+age, data=lung)
##
##             coef exp(coef) se(coef)     z       p
## sex     -0.567     0.567    0.168 -3.37 0.00075
## ph.ecog  0.470     1.600    0.113  4.16 3.1e-05
## age>70   0.307     1.359    0.187  1.64 0.10175
##
## Likelihood ratio test=31.6  on
## n= 227, number of events= 164

ggforest(fit)
```



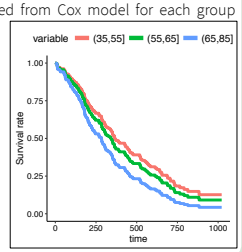
The function **ggcoxadjustedcurves()** from the **survminer** package plots Adjusted Survival Curves for Cox Proportional Hazards Model. Adjusted Survival Curves show how a selected factor influences survival estimated from a Cox model.
Note that these curves differ from Kaplan Meier estimates since they present expected ssurvival based on given Cox model.

```
library("survival")
library("survminer")

lung$sex <- ifelse(lung$sex == 1,
                   "Male", "Female")

fit <- coxph(Surv(time, status) ~
             sex + ph.ecog + age,
             data = lung)

ggcoxadjustedcurves(fit, data=lung,
             variable=lung$sex)
```



Note that it is not necessary to include the grouping factor in the Cox model. Survival curves are estimated from Cox model for each group defined by the factor independently.

```
lung$age3 <- cut(lung$age,
             c(35,55,65,85))

ggcoxadjustedcurves(fit, data=lung,
             variable=lung$age3)
```



This onepager presents the **survminer** package [Alboukadel Kassambara, Marcin Kosinski 2017] in version 0.3.