

Survival Analysis Using R

Antonio, Fabio Di Narzo

- Mon Model Based, Classic Statistical Inference and Regression Analysis
- Tue Nonparametric methods for Survival Analysis
- Wed Semi-parametric regression
- Thu Semi-parametric regression: model building and diagnostics
- Fri Penalized regression, Case Study

1 / 129

2 / 129

References

- ▶ Modeling Survival Data; Therneau, T., Grambsch, P.
- ▶ Modeling Survival Data: Extending the Cox Model. Springer-Verlag, 2000
- ▶ Survival Analysis: A Self-Learning Text, Third edition (Hardcover) by David G. Kleinbaum, Mitchel Klein
- ▶ Applied Survival Analysis Using R (Use R!) (Paperback) by Dirk F. Moore

Prerequisites

Classic Statistical Inference, Statistical Modeling

RStudio up and running

Being able to write and run an R script

Some extra R packages installed:

asaur, ggplot2, maxLik, plyr, reshape, survivalROC, glmnet, randomForestSRC

3 / 129

4 / 129

At the end of this course you should be able to perform statistical inference on survival data:

- ▶ estimate survival, parametrically or non-parametrically
- ▶ compare 2 or more groups
- ▶ make predictions

using the R statistical software

- ▶ Survival Analysis is the study of **survival times** and the factors that influence them
- ▶ Survival times, aka 'times to failure', have some distinguishing features:
 - ▶ non negative
 - ▶ the information is often only partially recorded (censoring)
- ▶ Aims:
 - ▶ Summarize and *interpret* survival/time to failure data
 - ▶ Make statistical *predictions*

Some Example Applications

The following examples are all taken from past students projects:

- ▶ Clinical trials : life expectancy of cancer patients by clinical outlook and treatment options
- ▶ Criminal Recidivism : risk of returning to prison by different follow-up policies
- ▶ Phone contract termination : risk by demographics and contract type
- ▶ Corruption in [country] : risk of corruption indictment by political party and region
- ▶ Unemployment Insurance claims : duration of unemployment by demographic and geographic factors
- ▶ Breast feeding behaviors : duration of breast feeding by ethnic, social and clinical background
- ▶ Roman Emperors reigns : risk of violent death by historical period, dynasty
- ▶ Heroin addiction : risk of relapse of heroin addicts by different treatment options
- ▶ Reliability of grid power lines : risk of failure by technology and geographic location

Case Study: overnight hospitalization

- ▶ hospitals are generally interested in minimizing the duration of patients hospitalization
- ▶ we have been hired by a small hospital as a new process manager. We want to use a data-driven approach possibly propose new policies to improve our performance w.r.t. duration of hospitalization
- ▶ after a good deal of poking the right people, finally a med student is forced to go through the medical records from the past few days
- ▶ We're handed back a small Excel file with the following columns:
 - DUR duration of hospitalization (days)
 - AGE (years)
 - SEX male/female
 - TEMP body temperature
 - WBC White blood cells per 100 ml of blood
 - ANTIB antibiotic use: yes/no
 - CULT blood culture taken: yes/no
 - SERV service type: medical/surgical
- ▶ **TASK:** Use the notebook to load the data into R, and get a sense of the data

- ▶ How many patients go through overnight hospitalization?
 - ▶ In our data, we find that 22 out of 25 patients go through overnight hospitalization. That is, 88% of the patients (95% CI: 0.69-0.97).
- ▶ Is the body temperature at admission predictive of the risk of longer hospitalization?
 - ▶ No
- ▶ What about blood works?
 - ▶ Neither
- ▶ Can we build a statistical model for the risk of being hospitalized overnight?
 - ▶ see next...

Y_i : subject i is hospitalized overnight

Y_i is a Random Variable, with a Bernoulli distribution:

$$Y_i \sim \text{Ber}(p_i)$$

Remember: $E[Y_i] = p_i$, which here we assume changes from subject to subject

Changes how?

$$\ln(p_i/(1 - p_i)) = \alpha + \beta x_i$$

where x_i is subject's i body temperature at admission

9 / 129

10 / 129

The Data Generating Process (cont.)

Briefly, here is our model:

$$(Y_i | X_i = x_i) \sim \text{Ber}(g(\alpha + \beta x_i)), \quad \text{i.i.d.}, \quad i = 1, \dots, n, \quad (\alpha, \beta) \in \mathbb{R}^2$$

- ▶ Note that x_i is part of the data. Our only parameters are α and β
- ▶ Note also that α and β **do not depend on i**

Likelihood function

What is the joint pdf of our data within the **sample space** of $n = 25$ samples?

$$\begin{aligned} P(Y_1 = y_1, \dots, Y_n = y_n; \alpha, \beta) &= \prod_{i=1}^n P(Y_i = y_i; \alpha, \beta) \\ &= \prod_{i=1}^n g(\alpha + \beta x_i)^{y_i} (1 - g(\alpha + \beta x_i))^{(1-y_i)} \end{aligned}$$

For our specific sample, the Y_i s are observed as $Y_i = y_i$. What's still unknown are the parameters α and β

The likelihood function:

$$L(\alpha, \beta; y_1, \dots, y_n) = \prod_{i=1}^n g(\alpha + \beta x_i)^{y_i} (1 - g(\alpha + \beta x_i))^{(1-y_i)}$$

$$(\widehat{\alpha}, \widehat{\beta})_{\text{ML}} = \max_{(\alpha, \beta) \in \mathbb{R}^2} L(\alpha, \beta; y_1, \dots, y_n)$$

Use R to:

- ▶ load and prepare the data
- ▶ write the likelihood function
- ▶ maximize it numerically
- ▶ answer the question: what's the impact of body temperature on the probability of an overnight hospitalization?
- ▶ solve the problem using canned logistic regression
- ▶ based on the model, predict the probability of overnight hospitalization for a new patient admitted with body temperature = 38°C

13 / 129

Probabilistic description of Duration Data

- ▶ Support: $0 \leq t < \infty$

The distribution can be specified through one of the following:

- ▶ **Survival function:**

$$S(t) = \Pr(T > t)$$

- ▶ **Cumulative Distribution Function (CDF):**

$$F(t) = \Pr(T \leq t) = 1 - \Pr(T > t) = 1 - S(t)$$

- ▶ **Probability Density Function (PDF):**

$$f(t) = F'(t) = -S'(t)$$

14 / 129

Hazard Function

A meaningful quantity linked to a survival distribution is the **Hazard function**:

$$h(t) = \lim_{\delta \rightarrow 0} \frac{\Pr(t < T < t + \delta | T > t)}{\delta}$$

and the derived **Cumulative Hazard function**:

$$H(t) = \int_0^t h(u) du$$

- ▶ Note:

$$h(t) = \frac{f(t)}{S(t)}$$

$$S(t) = \exp(-H(t))$$

15 / 129

16 / 129

- Mean survival:

$$\mu = E(T) = \int_0^\infty tf(t)dt = \int_0^\infty S(t)dt$$

Note: μ is only defined if $S(\inf) = 0$.

- Median survival

$$t : S(t) = 0.5$$

For $t \in [0, \infty)$:

- Survival function: $S(t) = \Pr(T > t)$, *right continuous*
- Cumulative Distribution Function (CDF): $F(t) = \Pr(T \leq t)$
- Probability Density Function (PDF): $f(t) = F'(t)$
- Hazard function: $h(t) = \lim_{\delta \rightarrow 0} \frac{\Pr(t < T < t + \delta | T > t)}{\delta}$
- Cumulative Hazard function: $H(t) = \int_0^t h(u)du$

17 / 129

18 / 129

Recap (2/2)

Some relationships allow us to switch from one quantity to another:

$$f(t) = F'(t)$$

$$h(t) = \frac{f(t)}{S(t)}$$

$$H(t) = \int_0^t h(u)du$$

$$S(t) = \exp(-H(t))$$

$$F(t) = 1 - S(t)$$

$$E(T) = \mu = \int_0^\infty tf(t)dt = \int_0^\infty S(t)dt$$

$$\text{median}(t) = \{t : S(t) = 0.5\} = S^{-1}(0.5)$$

Survival distribution: Exponential

- constant hazard: $h(t) = \lambda$
- cumulative hazard: $H(t) = \int_0^t \lambda du = \lambda \int_0^t du = \lambda t$
- mean: $\int_0^\infty e^{-\lambda t} dt = 1/\lambda$
- Exercise: can you determine:
 - Survival function
 - PDF
 - Median

19 / 129

20 / 129

- ▶ Survival function: $e^{-\lambda t}$
- ▶ PDF: $f(t) = -S'(t) = \lambda e^{-\lambda t}$
- ▶ Median: $0.5 = e^{-\lambda t} \implies t_{\text{med}} = \ln(2)/\lambda$

In R, the `rexp` function generates random samples from the exponential distribution

- ▶ Generate 100 samples from an exponential distribution with $\lambda = 0.5$
- ▶ Estimate from the simulated data:
 - ▶ mean
 - ▶ median
 - ▶ CDF (plot it)
 - ▶ Survival function (plot it)
 - ▶ PDF (plot it) (hint: `stats::density`)
 - ▶ bonus: hazard function (plot it)

How close are the values to their theoretical counterparts?

21 / 129

22 / 129

Exercise: solution (1/2)

```

1 y <- rexp(100, rate = 0.2)
2
3 mean(y)
4 1/0.2
5 median(y)
6 log(2) / 0.2
7
8 F <- ecdf(y)
9 plot(F)
10 curve(pexp(x, rate = 0.2), col = "red", add = TRUE)
11
12 S <- function(t) 1 - F(t)
13 curve(S(x), from = 0, to = 30)
14 curve(1 - pexp(x, rate = 0.2), col = "red", add = TRUE)

```

23 / 129

Exercise: solution (2/2)

```

1 f <- density(y, from = 0)
2 curve(dexp(x, rate = 0.2), col = "red", from = 0, to = 25)
3 lines(f)
4
5 h_empirical <- f$y / S(f$x)
6 plot(f$x, h_empirical, type = "l")
7 abline(h = 0.2, col = "red")

```

24 / 129

Distribution	RNG	parameters	mean
Exponential	rexp	rate	$1/\text{rate}$
Weibull	rweibull	shape=a, scale=b	$b\Gamma(1 + 1/a)$
Gamma	rgamma	shape=a, scale=b	$a \cdot b$
Log Normal	rlnorm	meanlog= μ , sdlog= σ	$e^{\mu+1/2\sigma^2}$

Exercise

- ▶ From each distribution, generate 100 random values, and estimate: mean, median, CDF, Survival Function, PDF, hazard

Use the following parameter values for data generation:

Weibull	a=0.5, b=2.5
Gamma	a=2.0, b=2.5
Log Normal	meanlog=1.2, sdlog=0.9

- ▶ Use the exponential distribution to model the hospitalization data
- ▶ Use linear regression
- ▶ Try more flexible parametric survival distributions

Case study: Novel Object Interaction (NOI) in rats

For each of 5 mice:

1. put it in the box and start the timer
2. visually follow the rat for 120s
3. take note of when the first interaction happens (physically touching the new object)

After running the experiment, our (precious) data table looks like this:

rat ID	time	status
rat1	55	0
rat2	50	1
rat3	70	1
rat4	120	0
rat5	110	1

where we adopted the convention:

$$\text{status} = \begin{cases} 0 & \text{the experiment was somehow interrupted: no interaction} \\ 1 & \text{an interaction actually happened at that time point} \end{cases}$$

Can we estimate the survival function from these data?

$$\hat{S}_7(t) = P(\widehat{T > t}) = ?$$

Say $t = 10$. What would be a reasonable estimate of $S(t)$?

$$0 \leq t < 50 : \quad \hat{S}_7(t) = 1.0$$

What happened at $t = 50$? One of the experimental subjects experienced the event. To be precise, 1 out of 5 participating subjects experienced the event:

$$t = 50 : \quad \hat{S}_7(t) = 1 - 1/5 = 4/5$$

Thinking about it, as no events happen between $t = 50$ and $t = 70$, we can actually write:

$$50 \leq t < 70 : \quad \hat{S}_7(t) = 4/5 = 0.8$$

We can start by ordering the table by time:

i	rat ID	time	status
1	rat2	50	1
2	rat1	55	0
3	rat3	70	1
4	rat5	110	1
5	rat4	120	0

$$\hat{S}_7(70) = ?$$

We can write:

$$\begin{aligned} P(T > 70) &= P((T > 70) \cap (T > 50)) \\ &= P(T > 70 | T > 50) \times P(T > 50) \end{aligned}$$

$$\hat{S}_7(70) = \left(1 - \frac{1}{3}\right) \times \frac{4}{5} \simeq 0.533$$

Again, as nothing happens between $t = 70$ and $t = 110$, we can write:

$$70 \leq t < 110 : \quad \hat{S}_7(t) = \frac{2}{3} \times \frac{4}{5} \simeq 0.533$$

similarly for the next event:

$$110 \leq t < 120 : \quad \hat{S}_7(t) = \frac{1}{2} \times \frac{2}{3} \times \frac{4}{5} \simeq 0.267$$

Let's add one more utility column to our sorted table:

i	rat ID	time	status	n
1	rat2	50	1	5
2	rat1	55	0	4
3	rat3	70	1	3
4	rat5	110	1	2
5	rat4	120	0	1

with n = number of subjects still under observation at that time point

29 / 129

30 / 129

Notation

We can summarize our calculations in this table:

i	j	rat ID	time	status	n	q	1-q	S
1	1	rat2	50	1	5	1/5	4/5	4/5 = 0.8
3	2	rat3	70	1	3	1/3	2/3	2/3 × 4/5 = 8/15 ≈ 0.533
4	3	rat5	110	1	2	1/2	1/2	1/2 × 8/15 = 4/15 ≈ 0.267

Congrats! We just re-discovered the **Kaplan-Meier estimator**

- ▶ T : time to failure
- ▶ U : time to censoring
- ▶ δ : $I[T < U]$
- ▶ observed data: $(\min(T, U), \delta)$

31 / 129

32 / 129

$$\hat{S}_{KM}(t) = \prod_{t_i \leq t} (1 - \hat{q}_i) = \prod_{t_i \leq t} \left(1 - \frac{d_i}{n_i}\right)$$

- ▶ n_i : # subjects at risk at time t_i
- ▶ d_i : # subjects failing at time t_i

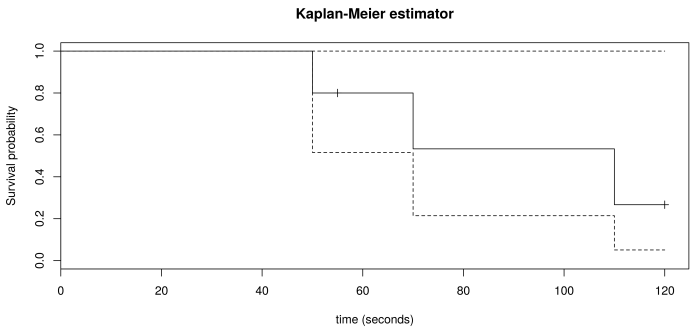
The KM estimator can be obtained with the `survival::survfit` function:

```
1 library(survival)
2 dat <- data.frame(ratID = paste0("rat", 1:5),
3                   time = c(55, 50, 70, 120, 110),
4                   status = c(0, 1, 1, 0, 1))
5 fit.KM <- survfit(Surv(time, failure) ~ 1, data = dat)
6 summary(fit.KM)
```

Call: `survfit(formula = Surv(time, status) ~ 1, data = dat)`

time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
50	5	1	0.800	0.179	0.5161	1
70	3	1	0.533	0.248	0.2142	1
110	2	1	0.267	0.226	0.0507	1

```
1 plot(fit.KM, mark.time = TRUE,
2      main = "Kaplan-Meier estimator",
3      ylab = "Survival probability",
4      xlab = "time (seconds)")
```



```
1 fit.KM
```

Call: `survfit(formula = Surv(time, status) ~ 1, data = dat)`

n	events	median	0.95LCL	0.95UCL
5	3	110	70	NA

Refined definition of median survival time:

- ▶ maximum time t such that $S(t) \geq 0.5$

- ▶ censoring is a main feature of survival data
- ▶ it happens when starting or ending events are not precisely observed
- ▶ in this course, we will focus on **right censoring**: the time to failure for some samples is only known to *exceed* a particular value
- ▶ censoring might happen because:
 - ▶ the event of interest did not happen by the end of the study
 - ▶ e.g., we turn on 100 lightbulbs for 30 days, and we record burnout times; for lightbulbs still on after 30 days, we can only say that the *survival time* was > 30 days
 - ▶ the sample drops out from the study from unrelated causes
 - ▶ e.g., in a clinical trial, 200 subjects might be administered a new drug and their prognosis followed for 10 years; in these 10 years, some of the 200 subjects might move to a different city, die of unrelated causes, or just plain decide to stop participating in the study

Back to our data table:

i	j	rat ID	time	status	n	q
1	1	rat2	50	1	5	1/5
3	2	rat3	70	1	3	1/3
4	3	rat5	110	1	2	1/2

The q_i s can be seen as empirical estimates of instantaneous risks at the times t_j . They can be *cumulated*, to get a corresponding empirical cumulated risk:

$$\hat{H}_j = \sum_{i=1}^j q_i$$

From those, an estimator of survival could be:

$$\hat{S}_7(t_j) = e^{-\hat{H}_j}$$

An alternative approach (cont.)

Nelson-AAalen estimator: definition

Here is how the calculation looks like:

i	j	rat ID	time	status	n	q	H	S
1	1	rat2	50	1	5	1/5	1/5	$e^{-1/5} \simeq 0.819$
3	2	rat3	70	1	3	1/3	8/15	$e^{-8/15} \simeq 0.587$
4	3	rat5	110	1	2	1/2	31/30	$e^{-31/30} \simeq 0.356$

We just computed the **Nelson-AAalen estimator** of survival

- ▶ AKA Fleming-Harrington estimator
- ▶ based on the relationship between cumulative hazard and survival function

$$\hat{H}_{NA}(t) = \sum_{t_j \leq t} \frac{d_i}{n_i}$$

$$\hat{S}_{NA}(t) = e^{-\hat{H}_{NA}(t)}$$

```
1 fit.NA <- survfit(Surv(time, status) ~ 1, data = dat, type = "fh")
2 summary(fit.NA)
```

Call: survfit(formula = Surv(time, status) ~ 1, data = dat, type = "fh")

time	n.risk	n.event	survival	std.err	lower	95% CI upper	95% CI
50	5	1	0.819	0.183	0.5282		1
70	3	1	0.587	0.273	0.2356		1
110	2	1	0.356	0.301	0.0677		1

Case study: questions

- ▶ Express the Progress-Free Survival (PFS) times in *months*
- ▶ Estimate and plot the survival function using the KM and NA methods
- ▶ What's the median survival (and CI) according to the two methods?

Cancer Chemother Pharmacol (2014) 73:155–161
DOI 10.1007/s00280-014-2449-1

ORIGINAL ARTICLE

A phase II trial of Xeloda and oxaliplatin (XELOX) neo-adjuvant chemotherapy followed by surgery for advanced gastric cancer patients with para-aortic lymph node metastasis

Yan Wang · Yi-yi Yu · Wei Li · Yi Feng · Jun Hou ·
Yuan Ji · Yi-hong Sun · Kun-tang Shen ·
Zhao-shu Shen · Xiang-Qin · Tian-shu Liu

Received: 8 January 2014 / Accepted: 11 March 2014 / Published online: 21 April 2014
© The Author(s) 2014. This article is published with open access at Springerlink.com

Purpose Gastric cancer with para-aortic lymph node (PAN) involvement is regarded as advanced disease, and only chemotherapy is recommended from the guidelines. In unresectable cases, neoadjuvant chemotherapy could prolong survival if conversion to resectability could be achieved.

Methods The study was a single-arm phase II trial. Patients who were diagnosed with gastric cancer and PAN involvement (Stations No. 16a2/16b1) were treated with capecitabine and oxaliplatin combination chemotherapy every 3 weeks for a maximum of six cycles. After every two cycles, abdominal computed tomographic scans were repeated to evaluate the response, and surgery was performed at the physician's discretion in patients with sufficient tumor response, followed by chemotherapy with the same regimen to complete a total of six cycles. The primary end point was the response rate of the preoperative chemotherapy. The secondary end points were R0 resection rate, progression-free survival (PFS), overall survival (OS), and adverse events.

To load the PFS data in R:

```
1 library(asaur)
2 dat <- gastricXelox
```

Case study: R code

```
1 library(survival)
2 library(asaur)
3
4 dat <- gastricXelox
5 dat$months <- with(dat, timeWeeks * 7 / 30.5)
6 dat$S <- with(dat, Surv(months, delta))
7 fit.KM <- survfit(S ~ 1, data = dat, type = "kaplan-meier",
8                   conf.type = "log-log")
9 fit.NA <- survfit(S ~ 1, data = dat, type = "fleming-harrington",
10                  conf.type = "log-log")
11
12 plot(fit.KM)
13 plot(fit.NA)
```

A quality metric for a trial is the *median follow up time*:

```
1 dat$delta.followup <- 1 - dat$delta
2 survfit(Surv(months, delta.followup) ~ 1, type = "k",
3         conf.type = "log-log")
```

n	events	median	0.95LCL	0.95UCL
48.0	16.0	27.5	13.5	42.9

Back to our 5 rats:

rat ID	time	status	group
rat1	55	0	0
rat2	50	1	1
rat3	70	1	0
rat4	120	0	1
rat5	110	1	1

As it turns out, they were belonging to 2 different experimental groups: **group 1**, which was sleep deprived, and **group 0**, which followed a natural sleep pattern.

Is there evidence of a different stress level between the two (precious, though tiny) groups?

45 / 129

Comparing Survival between 2 samples

Null hypothesis:

$$H_0 : S_1(t) = S_0(t)$$

- $S_1(t)$: Survival Distribution in group 1 (e.g. *treated*)
- $S_0(t)$: Survival Distribution in group 0 (e.g. *control*)

A minor note: the Lehman alternative:

$$H_A : S_1(t) = [S_0(t)]^\psi$$

or, equivalently:

$$h_1(t) = \psi h_0(t)$$

that is, the hazard functions of the two groups are proportional, with $H_0 : \psi = 1$ vs $H_A : \psi \neq 1$



46 / 129

For each failure time t_i , we build the following table:

	Control	Treatment	Total
Failures	d_{0i}	d_{1i}	d_i
Non-failures	$n_{0i} - d_{0i}$	$n_{1i} - d_{1i}$	$n_i - d_i$
Total	n_{0i}	n_{1i}	n_i

Under the assumption of independence of the two groups, conditional on the margins, d_{0i} follows the hypergeometric distribution:

$$E(d_{0i} | n_i, d_i, n_{0i}, n_{1i}) = n_{0i} d_i / n_i$$

$$\text{Var}(d_{0i} | n_i, d_i, n_{0i}, n_{1i}) = \frac{n_{0i} n_{1i} d_i (n_i - d_i)}{n_i^2 (n_i - 1)}$$

47 / 129

48 / 129

The logrank test (cont.)

Summing over all time points t_i :

$$U_0 = \sum_i (d_{0i} - e_{0i})$$

with variance:

$$\text{Var}(U_0) = \sum_i \text{Var}(d_{0i}) = V_0$$

Finally, the logrank test:

$$\frac{U_0^2}{V_0} \sim \chi_1^2$$

The logrank test in R

Using the `survival::survdif` function:

```
1 dat <- data.frame(ratID = paste0("rat", 1:5),
2                   time = c(55, 50, 70, 120, 110),
3                   status = c(0, 1, 1, 0, 1),
4                   group = c(0, 1, 0, 1, 1))
5
6 survdiff(Surv(time, status) ~ group, data = dat)
```

Call:

```
survdif(formula = Surv(time, status) ~ group, data = dat)
```

	N	Observed	Expected	(O-E) ² /E	(O-E) ² /V
group=0	2	1	0.733	0.0970	0.154
group=1	3	2	2.267	0.0314	0.154

Chisq= 0.2 on 1 degrees of freedom, p= 0.7

49 / 129

50 / 129

The Fleming-Harrington test

A weighted variation on the logrank test:

$$U_0(w) = \sum w_i (d_{0i} - e_{0i})$$

$$\text{Var}(U_0) = \sum w_i^2 v_{0i} = V_0(w)$$

with:

$$w_i = N(\hat{S}_{KM}(t_i))^\rho$$

- ▶ $\rho = 0$: logrank test
- ▶ $\rho = 1$: aka Prentice modification of the Gehan-Wilcoxon test: higher weights on *earlier* survival times

Case study: the pancreatic dataset

```
1 library(asaur)
2 dat <- pancreatic
3 head(dat)
```

	stage	onstudy	progression	death
1	M	12/16/2005	2/2/2006	10/19/2006
2	M	1/6/2006	2/26/2006	4/19/2006
3	LA	2/3/2006	8/2/2006	1/19/2007
4	M	3/30/2006	.	5/11/2006
5	LA	4/27/2006	3/11/2007	5/29/2007
6	M	5/7/2006	6/25/2006	10/11/2006

51 / 129

52 / 129

Case study: preparing the data for analysis

```

1 fmt <- "%m/%d/%Y"
2 dat <- within(dat, {
3   onstudy <- as.Date(as.character(onstudy), format = fmt)
4   progression <- as.Date(as.character(progression), format = fmt)
5   death <- as.Date(as.character(death), format = fmt)
6   OS <- death - onstudy
7   PFS <- pmin(progression - onstudy, OS)
8   PFS[is.na(PFS)] <- OS[is.na(PFS)]
9   PFS <- Surv(as.numeric(PFS / 30.5))
10  OS <- Surv(as.numeric(OS / 30.5))
11 })

```

53 / 129

Case study: comparing survival by stage

```

1 survdiff(PFS ~ stage, data = dat)

```

	N	Observed	Expected	(O-E) ² /E	(O-E) ² /V
stage=LA	8	8	12.3	1.49	2.25
stage=M	33	33	28.7	0.64	2.25

Chisq= 2.2 on 1 degrees of freedom, p= 0.134

```

1 survdiff(PFS ~ stage, data = dat, rho = 1)

```

	N	Observed	Expected	(O-E) ² /E	(O-E) ² /V
stage=LA	8	2.34	5.88	2.128	4.71
stage=M	33	18.76	15.22	0.822	4.71

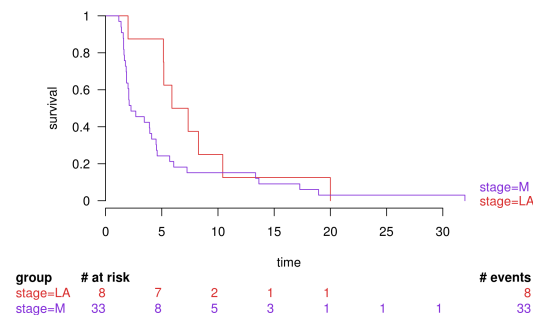
Chisq= 4.7 on 1 degrees of freedom, p= 0.0299

Case study: estimating survival by stage

```

1 surv.KM <- survfit(PFS ~ stage, data = dat)
2 plot(surv.KM)

```



54 / 129

Exercises

- ▶ What's the median *Overall Survival* of a patient with Locally Advanced (LA) pancreatic cancer? And that of a patient with Metastatic (M) cancer?
- ▶ Can you provide confidence intervals for your estimates?
- ▶ Do the two stages experience significantly different survival?
- ▶ What's the probability (and 95% CI) of surviving more than a year within each group?

55 / 129

56 / 129

- ▶ median OS + CIs: fit formula $OS \sim stage$, then `summary(fit)`
- ▶ plot the curves for qualitative assessment, `survdiff(OS ~ stage)` for logrank test
- ▶ `summary(fit, time = 12)` will give survival and CIs at 12 months

- ▶ Sometimes we want to compare survival between 2 groups *controlling* for potentially confounding factors, e.g.:
 - ▶ gender
 - ▶ age group
 - ▶ hospital
 - ▶ ...
- ▶ When this factor is categorical, we can use a **stratified logrank test**

$$\chi^2 = \frac{\left(\sum_{g=1}^G U_{0g}\right)^2}{\sum_{g=1}^G V_{0g}^2}$$

distributed as a χ_1^2

57 / 129

58 / 129

Case study: the pharmacoSmoking dataset

```

1 dat <- pharmacoSmoking
2
3 survdiff(Surv(ttr, relapse) ~ grp, data = dat)
4 #               N Observed Expected (O-E)^2/E (O-E)^2/V
5 #grp=combination 61      37      49.9      3.36      8.03
6 #grp=patchOnly   64      52      39.1      4.29      8.03
7 #
8 # Chisq= 8  on 1 degrees of freedom, p= 0.00461
9
10 table(dat$AgeGroup2)
11 #21-49  50+
12 #   66   59
13
14 survdiff(Surv(ttr, relapse) ~ grp + strata(ageGroup2), data = dat)
15 #               N Observed Expected (O-E)^2/E (O-E)^2/V
16 #grp=combination 61      37      49.1      2.99      7.03
17 #grp=patchOnly   64      52      39.9      3.68      7.03
18 #
19 # Chisq= 7  on 1 degrees of freedom, p= 0.008

```

Exercises

- ▶ Assess the significance of the treatment stratifying by employment status
- ▶ Can you estimate survival in the 4 groups:
 - ▶ grp=combination/pathOnly x employment=ft/pt
- ▶ Assess the efficacy of the treatment combination therapy separately within patients working full time (*ft*) and part-time (*pt*)

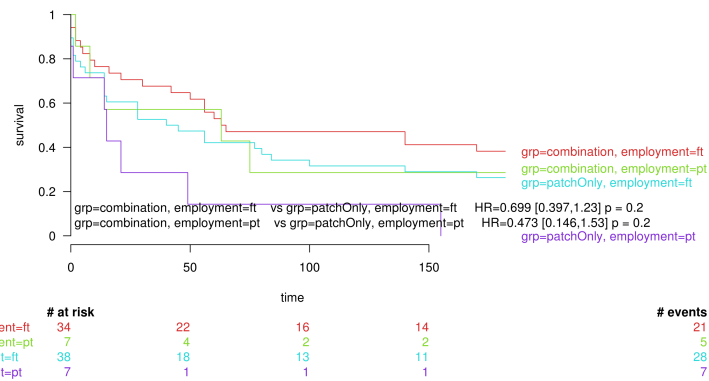
59 / 129

60 / 129

```

1 survdiff(Surv(ttr, relapse) ~ grp + strata(employment), data = dat
2 )
3
4 #           N Observed Expected (O-E)^2/E (O-E)^2/V
5 #grp=combination 61      37      50.3      3.50      8.58
6 #grp=patchOnly   64      52      38.7      4.54      8.58
7 #
8 # Chisq= 8.6  on 1 degrees of freedom , p= 0.00339

```



61 / 129

- ▶ 1 sample inference: KM, HF (`survival::survfit`)
- ▶ 2 samples comparison: logrank test + weighted variations (`survival::survdif`)

62 / 129

Proportional hazards model

- ▶ We saw methods for comparing 2 groups
- ▶ A more general approach is needed for comparing multiple groups, assessing the effect of continuous factor, and, in general, performing regression analysis
- ▶ Meet the Cox Proportional Hazards model:

$$h_i(t) = h_0(t) \exp(\mathbf{x}_i' \boldsymbol{\beta})$$

- ▶ $h_i(t)$: hazard for subject i at time t
- ▶ h_0 : baseline hazard function
- ▶ \mathbf{x}_i : vector of covariates for subject i
- ▶ $\boldsymbol{\beta}$: vector of effects of each covariate on risk

63 / 129

Proportional hazards model (cont.)

- ▶ Given an observed dataset $\{t_i, \delta_i, \mathbf{x}_i : i = 1, \dots, n\}$, one can estimate $\boldsymbol{\beta}$ without having to specify the baseline hazard h_0
- ▶ The CPH model is thus called *semi-parametric*
- ▶ As failure times are generally *censored*, we cannot compute the classic likelihood, but rather the so called *Partial Likelihood*, which properly takes into account censoring times similarly to how it's done in the KM estimator
- ▶ We'll call $\hat{\boldsymbol{\beta}}$ the estimator which maximizes the Partial Likelihood for a given dataset

64 / 129

- ▶ Failure times $t_j : t_1 \leq t_2 \leq \dots \leq t_j \leq \dots \leq t_D$
- ▶ At time t_j , subject $i(j)$ fails, with hazard $h_i(t_j) = h_0(t_j)\exp(\mathbf{x}_{i(j)}\boldsymbol{\beta})$
- ▶ At failure time t_j , R_j subjects at risk
- ▶ Partial Likelihood:

$$l(\boldsymbol{\beta}) = \prod_{j=1}^D \frac{h_0(t_j)\exp(\mathbf{x}_{i(j)}\boldsymbol{\beta})}{\sum_{k \in R_j} h_0(t_j)\exp(\mathbf{x}_k\boldsymbol{\beta})}$$

$$= \prod_{j=1}^D \frac{\exp(\mathbf{x}_{i(j)}\boldsymbol{\beta})}{\sum_{k \in R_j} \exp(\mathbf{x}_k\boldsymbol{\beta})}$$

Our beloved rats:

rat ID	time	status	group
rat1	55	0	0
rat2	50	1	1
rat3	70	1	0
rat4	120	0	1
rat5	110	1	1

In R:

```
1 dat <- data.frame(ratID = paste0("rat", 1:5),
2                   time = c(55, 50, 70, 120, 110),
3                   failure = c(0, 1, 1, 0, 1),
4                   group = c(0, 1, 0, 1, 1))
```

65 / 129

66 / 129

Comparing 2 groups (cont.)

```
1 library(survival)
2 fit <- coxph(Surv(time, failure) ~ group, data = d)
3 summary(fit)
```

Comparing 2 groups (cont.)

```
1 coxph(formula = Surv(time, failure) ~ x, data = dat)
2
3 n= 5, number of events= 3
4
5      coef exp(coef) se(coef)      z Pr(>|z|)
6 x -0.5493    0.5774    1.4179 -0.387    0.698
7
8      exp(coef) exp(-coef) lower .95 upper .95
9 x    0.5774      1.732    0.03585    9.297
10
11 Concordance= 0.5   (se = 0.202 )
12 Rsquare= 0.029   (max possible= 0.743 )
13 Likelihood ratio test= 0.15 on 1 df,  p=0.7
14 Wald test         = 0.15 on 1 df,  p=0.7
15 Score (logrank) test = 0.15 on 1 df,  p=0.7
```

67 / 129

68 / 129

$$h_i(t) = h_0(t)\exp(x_i\beta)$$

$$x_i = \begin{cases} 0 & \text{sample } i \text{ is a control} \\ 1 & \text{sample } i \text{ is treated} \end{cases}$$

$$\hat{\beta} = -0.549 \pm 1.418 \times 1.96$$

What's the risk of a sleep deprived (treated) rat compared to a control?

$$\begin{aligned} \frac{h_1(t)}{h_0(t)} &= \frac{h_0(t)\exp(1 \times \hat{\beta})}{h_0(t)\exp(0 \times \hat{\beta})} = \exp((1 - 0)\hat{\beta}) \\ &= \exp(\hat{\beta}) = 0.577 \end{aligned}$$

- ▶ in general, $\exp(\beta)$ is the hazard ratio associated with one unit increase of the regressor
- ▶ for 0/1 binary variables, it is e.g. a comparison between the group $x = 1$ and the group $x = 0$ (treated vs control, male vs female, etc.)
- ▶ more generally, x can be *continuous* (e.g., age of the subject)

69 / 129

Continuous covariates

```
1 d <- data.frame(patient = 1:6,
2                 time = c(6, 7, 10, 15, 19, 25),
3                 censored = c(1, 0, 1, 1, 0, 1),
4                 age = c(67, 62, 34, 41, 46, 28))
5 fit <- coxph(Surv(time, censored) ~ age, data = d)
```

Questions:

- ▶ is the effect of age on risk significant?
- ▶ what's the HR for a 1 year increase of age?
- ▶ what's the HR for a 10 years increase of age?

70 / 129

Continuous covariates (cont.)

```
1 n= 6, number of events= 4
2
3      coef exp(coef) se(coef)      z Pr(>|z|)
4 age 0.07606   1.07903  0.07316  1.04  0.298
5
6      exp(coef) exp(-coef) lower .95 upper .95
7 age      1.079      0.9268   0.9349   1.245
8
9 Concordance= 0.7   (se = 0.22 )
10 Rsquare= 0.209   (max possible= 0.76 )
11 Likelihood ratio test= 1.41 on 1 df,  p=0.2356
12 Wald test           = 1.08 on 1 df,  p=0.2985
13 Score (logrank) test = 1.33 on 1 df,  p=0.2482
```

```
1 exp(0.076 * 10)
2 # [1] 2.138276
```

71 / 129

72 / 129

Multiple covariates

```
1 library( asaur )
2 dat <- pharmacoSmoking
3 names( pharmacoSmoking )

[1] "id"      "ttr"      "relapse"  "grp"
[5] "age"      "gender"    "race"
   "employment"
[9] "yearsSmoking" "levelSmoking" "ageGroup2" "ageGroup4"
[13] "priorAttempts" "longestNoSmoke"
```

Multiple covariates (cont.)

```
1 n= 125, number of events= 89
2
3      coef exp( coef) se( coef) z Pr(>|z|)
4 grppatchOnly 0.5656340 1.7605636 0.2181634 2.593 0.00952 **
5 age -0.0220948 0.9781475 0.0097572 -2.264 0.02355 *
6 genderMale -0.1215514 0.8855455 0.2334349 -0.521 0.60257
7 priorAttempts 0.0002078 1.0002079 0.0010898 0.191 0.84876
8
9
10      exp( coef) exp(- coef) lower .95 upper .95
11 grppatchOnly 1.7606 0.5680 1.1480 2.700
12 age 0.9781 1.0223 0.9596 0.997
13 genderMale 0.8855 1.1292 0.5604 1.399
14 priorAttempts 1.0002 0.9998 0.9981 1.002
15
16 Concordance= 0.623 ( se = 0.034 )
17 Rsquare= 0.107 ( max possible= 0.998 )
```

Multiple covariates (cont.)

```
1 library( survival )
2 fit <- coxph( Surv( ttr , relapse ) ~ grp + age + gender +
3               priorAttempts ,
4               data = dat )
5 summary( fit )
```

Multiple covariates: interpretation

What's the risk of relapse in subjects treated with patch only, compared to subjects with combination therapy, *all other covariates being the same*?

$$\begin{aligned} & \frac{h(t|grpPO = 1, age = X, genderMale = Y, priorAttempts = Z)}{h(t|grpPO = 0, age = X, genderMale = Y, priorAttempts = Z)} \\ &= \frac{\exp(1\beta_1 + X\beta_2 + Y\beta_3 + Z\beta_4)}{\exp(0\beta_1 + X\beta_2 + Y\beta_3 + Z\beta_4)} \\ &= \exp[(1 - 0)\beta_1 + (X - X)\beta_2 + (Y - Y)\beta_3 + (Z - Z)\beta_4] \\ &= \exp(\beta_1) \\ &= \exp(0.5656) = 1.7606 \end{aligned} \tag{1}$$

```

1 dat$grp <- relevel(dat$grp, ref = "patchOnly")
2 update(fit)

```

	coef	exp(coef)	se(coef)	z	p
grpcombination	-0.565634	0.568000	0.218163	-2.59	0.0095
age	-0.022095	0.978147	0.009757	-2.26	0.0235
genderMale	-0.121551	0.885546	0.233435	-0.52	0.6026
priorAttempts	0.000208	1.000208	0.001090	0.19	0.8488

$$\hat{h}_0(t_i) = d_i / \sum_{j \in R_i} \exp(x_j \hat{\beta})$$

$$\hat{H}_0(t) = \sum h_0(t_j), \quad t_j \leq t$$

$$\hat{S}_0(t) = \exp(-\hat{H}_0(t))$$

$$\hat{S}(t|x) = [S_0(t)]^{\exp(x\hat{\beta})}$$

Cfr. R function `survival::survfit.coxph`

77 / 129

78 / 129

Predicting Survival: Exercise

```

1 d <- data.frame(patient = 1:6,
2                 time = c(6, 7, 10, 15, 19, 25),
3                 censored = c(1, 0, 1, 1, 0, 1),
4                 age = c(67, 62, 34, 41, 46, 28))

```

Predict and plot survival curves at age 20, 50 and 70

Predicting Survival: Exercise (cont.)

```

1 fit <- coxph(Surv(time, censored) ~ age, data = d)
2 pred <- survfit(fit, newdata = data.frame(age = c(20, 40, 60)))
3 plot(pred, col = 1:3)

```

79 / 129

80 / 129

- ▶ when no censoring, classical statistical methods (MLE and friends)
- ▶ with right-censored data:
 - ▶ 1 sample inference: KM, HF (`survival::survfit`)
 - ▶ 2 samples comparison: logrank test + weighted variations (`survival::survdif`)
 - ▶ continuous factors and/or multiple covariates: Cox regression (`survival::coxph`, `survival::survfit`)

- ▶ comparing nested models
- ▶ comparing non-nested models
- ▶ assessing goodness of fit
- ▶ checking model assumptions

81 / 129

Comparing models

We will consider the following models for the `pharmacoSmoking` dataset:

- ▶ M0: *no covariates* (hint: ~ 1)
- ▶ MA: `ageGroup4`
- ▶ MB: `employment`
- ▶ MC: `ageGroup4 + employment`

Both models MA and MB are nested into model MC, however MA and MB are not nested into eachother.

Exercise: fit the 3 models in R and store them in the variables M0, MA, MB and MC. We will be comparing these models

82 / 129

An aside: the Likelihood Ratio Test

$$Y_i \sim f(\theta), \quad i = 1, \dots, n \quad \theta \in \Theta$$

$$H_0 : \theta \in \Theta_0$$

$$H_1 : \theta \in \Theta_1$$

A very important test statistic for H_0 is the LRT:

$$\text{LRT}_n = -2 \ln \frac{\sup\{L(\theta; \mathbf{y}) : \theta \in \Theta_0\}}{\sup\{L(\theta; \mathbf{y}) : \theta \in \Theta\}}$$

83 / 129

84 / 129

Under H_0 , $n \rightarrow \infty$:

$$\text{LRT}_n \xrightarrow{p_{n \rightarrow \infty}} \chi_p^2$$

where p is the difference in dimensionality between Θ_0 and Θ .

Note: a necessary condition for the Theorem to hold is that Θ_0 is in the interior of Θ (i.e., Θ_0 should not be on the boundaries of Θ).

```
1 anova(MA, MC)
```

```
1 Analysis of Deviance Table
2 Cox model: response is Surv(ttr, relapse)
3 Model 1: ~ ageGroup4
4 Model 2: ~ ageGroup4 + employment
5      loglik   Chisq Df P(>|Chi|)
6 1 -380.04
7 2 -377.76 4.5666 2 0.1019
```

$H = (\beta_1, \beta_2, \beta_3, \beta_4, \beta_5) \in \mathbb{R}^5$: Model 2.
 $H_0 = (\beta_4 = 0, \beta_5 = 0) \in \mathbb{R}^3$: Model 1.

we cannot reject H_0

85 / 129

86 / 129

Non-nested models: AIC

Step-wise model selection based on AIC

$$\text{AIC} = -2\log\text{Lik}(\hat{\beta}) + 2 \cdot k \quad (2)$$

The *smaller* the *better*

```
1 fits <- list(M0 = M0, MA = MA, MB = MB, MC = MC)
2 sapply(fits, AIC)
3 ##      MA      MB      MC
4 ## 766.0860 774.2464 765.5194
```

```
1 Mfull <- coxph(Surv(ttr, relapse) ~ grp + gender + race +
2               employment + yearsSmoking + levelSmoking +
3               ageGroup4 + priorAttempts + longestNoSmoke,
4               data = dat)
5 MAIC <- step(Mfull)
```

87 / 129

88 / 129

First step:

```

1 Start:  AIC=770.2
2 Surv(ttr, relapse) ~ grp + gender + race + employment +
3   yearsSmoking +
4   levelSmoking + ageGroup4 + priorAttempts + longestNoSmoke
5
6      Df    AIC
7 - race      3 766.98
8 - yearsSmoking 1 768.20
9 - gender      1 768.20
10 - priorAttempts 1 768.24
11 - levelSmoking 1 768.47
12 - longestNoSmoke 1 769.04
13 <none>      770.20
14 - employment  2 772.45
15 - ageGroup4    3 774.11
    - grp         1 776.80

```

Check ?step for further options

- ▶ Harrell's Concordance Index: fraction of pairs of patients whose survival times are correctly ordered by the model-fitted hazard
- ▶ the higher, the better
- ▶ in R, output of `summary(fit.coxph)`

89 / 129

Predictive power: AUC

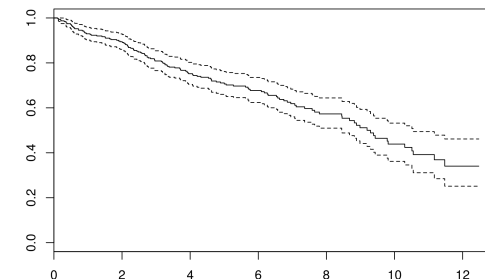
- ▶ A common measure of predictive power are the *ROC curve* (False Positive rate vs True Positive rate) and the associated *AUC*
- ▶ Their computation for survival data is complicated by the presence of censoring
- ▶ One can however estimate *time-dependent ROC curves* via Kaplan-Meier or Nearest Neighbor methods of Heagerty, Lumley & Pepe (Biometrics, Vol 56 No 2, 2000, PP 337-344)
- ▶ Conveniently implemented in the `survivalROC` R package

AUC (cont.)

```

1 library(survival)
2 library(survivalROC)
3
4 data(mayo)
5 plot(survfit(Surv(time / 365.25, censor) ~ 1, data = mayo))

```



91 / 129

90 / 129

92 / 129

AUC (cont.)

```

1 ROC.4 <- survivalROC (Stime = mayo$time,
2                       status = mayo$censor,
3                       marker = mayo$mayoscore4,
4                       predict.time = 365.25 * 5,
5                       method="KM")
6
7 ROC.5 <- survivalROC (Stime = mayo$time,
8                       status = mayo$censor,
9                       marker = mayo$mayoscore5,
10                      predict.time = 365.25,
11                      method = "KM")
12
13 ROC <- list (mayo4 = ROC.4, mayo5 = ROC.5)
14
15 sapply (ROC, " [ ", "AUC")
16 ##      mayo4      mayo5
17 ## 0.8257006 0.9180251

```

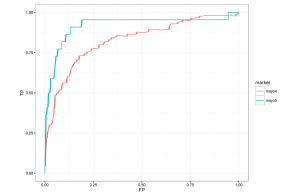
AUC (cont.)

We can plot the ROC curves using e.g. ggplot:

```

1 dfl <- lapply (ROC, function(x) with(x, data.frame(FP, TP)))
2 for (nm in names(dfl)) {
3   dfl[[nm]]$marker <- nm
4 }
5 dat <- do.call(rbind, dfl)
6
7 library(ggplot2)
8 ggplot(dat, aes(FP, TP, color = marker)) +
9   geom_line() +
10  theme_bw(base_size = 9)

```



93 / 129

94 / 129

AUC: exercise

Lets select a cutoff for mayoscore5 with FP = 10%:

```

1 cutoff <- with(ROC$mayo5, min(cut.values[FP <= 0.1]))
2 ## 7.511961

```

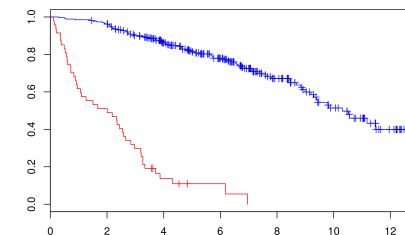
Question: can you compare the survival of patients with low vs high risk according to our chosen cutoff of mayoscore5?

AUC: solution

```

1 mayo$prediction <- ifelse(mayo$mayoscore5 <=
2 cutoff, "low_risk", "high_risk")
3
4 plot(survfit(Surv(time/365, censor) ~ prediction, data = mayo),
5      col = c("red", "blue"))

```



95 / 129

96 / 129

- ▶ A model was built and estimated, but how well are we fitting the data?
- ▶ In linear regression, we can look at patterns in the model *residuals* (observed value – model prediction)
- ▶ For Cox regression, we have *martingale residuals*
 - ▶ they sum to 1
 - ▶ each is distributed between $-\infty$ and $+1$
 - ▶ each has an expected value of 0
 - ▶ their sum of squares **is not** an indicator of goodness of fit
 - ▶ patterns might suggest alternative functional forms for continuous covariates
- ▶ In R, we use `residuals(fit, type = 'martingale')`, from the `survival` package

```

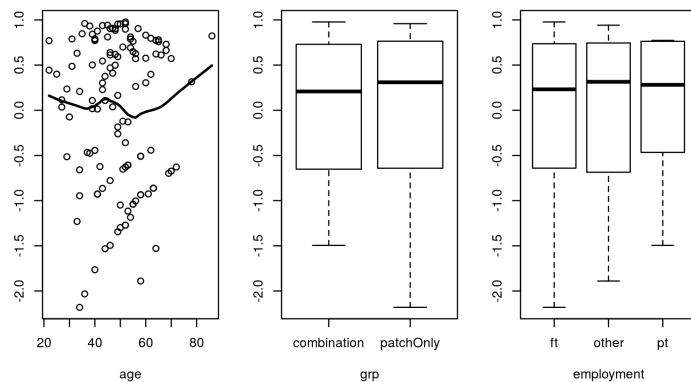
1 library(survival)
2 library(asaur) ## dataset
3
4 data(pharmacoSmoking)
5 dat <- pharmacoSmoking
6 fit <- coxph(Surv(ttr, relapse) ~ grp + age + employment, data =
7   dat)
8 dat$residual <- residuals(fit, type = 'martingale')
9
10 with(dat, {
11   plot(age, residual)
12   lines(lowess(age, residual), lwd = 2)
13
14   plot(residual ~ grp)
15
16   plot(residual ~ employment)
17
18 })

```

97 / 129

98 / 129

Martingale Residuals in R (cont.)



Case deletion residuals

- ▶ some samples might have a large impact on the final estimates
- ▶ we don't like it, as possibly all of our results (in extreme cases) might be driven by a single sample!
- ▶ such influential samples can be identified by estimating β twice: once with all the samples, and once without a specific sample i , and measuring the difference in β
- ▶ in R, `residuals(fit, type = 'dfbetas')`

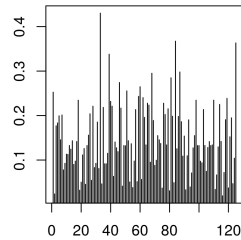
99 / 129

100 / 129

```

1 dfbetas <- residuals(fit, type = 'dfbetas')
2 dat$dfbetas <- sqrt(rowSums(dfbetas^2))
3
4 plot(dat$dfbetas, type = 'h')
5 abline(h = 0)

```



- ▶ one key assumption of the Cox model is the proportionality of hazards
- ▶ if we are comparing 2 groups:

$$S_1(t) = [S_0(t)]^{\exp(\beta)}$$

- ▶ by taking the log of both sides:

$$\log(S_1(t)) = \exp(\beta) \cdot \log[S_0(t)]$$

- ▶ finally, we can negate both sides and take a logarithm again:

$$\log(-\log(S_1(t))) = \beta + \log(-\log(S_0(t)))$$

- ▶ in this scale ($g(u) = \log(-\log(u))$), S_0 and S_1 should be **parallel**

101 / 129

102 / 129

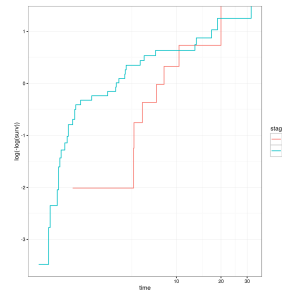
Proportionality of hazards: complementary log-log plot

Recall the 'pancreatic' dataset from the logrank test chapter:

```

1 library(plyr)
2 dat <- pancreatic
3 surv <- ddply(dat, .(stage),
4   function(x) {
5     fit <- survfit(PFS ~ 1, data =
6       x)
7     data.frame(time = fit$time,
8       surv = fit$surv)
9   })
10 ggplot(surv,
11   aes(x = time,
12     y = log(-log(surv)),
13     color = stage)) +
14   geom_step() +
15   coord_trans(x = "log")

```



Schoenfeld Residuals

```

1 dat <- pancreatic
2 residual.sch <- cox.zph(fit)
3
4 fit <- coxph(PFS ~ stage, data = dat)
5 residual.sch <- cox.zph(fit)
6 ##           rho chisq      p
7 ## stageM -0.328  3.86 0.0496
8
9 plot(residual.sch)

```

103 / 129

104 / 129

- ▶ does it really matter?
- ▶ stratification
- ▶ truncation

- ▶ Remember the Cox model:

$$h_i(t) = h_0(t) \exp(\mathbf{x}_i \beta) \quad \forall i$$

where *all samples* share the same baseline hazard $h_0(t)$

- ▶ We can somewhat relax this assumption, and allow for 2 (or more) separate baseline hazards in different **strata** of the samples
- ▶ Stratified Cox model, strata A and B :

$$h_i(t) = \begin{cases} h_A(t) \exp(\mathbf{x}_i \beta) & i \in A \\ h_B(t) \exp(\mathbf{x}_i \beta) & i \in B \end{cases}$$

- ▶ **Question:** how is this different from just modeling A and B separately?
- ▶ Analyze the `asauro::pharmacoSmoking` dataset, stratifying by employment type

105 / 129

106 / 129

Truncation

- ▶ Proportionality of hazards might hold for a shorter, initial time span
- ▶ If so, we can restrict the analysis to a properly defined, initial time period
- ▶ How, in practice?
- ▶ Introduce a new, **truncated** time variable:

$$t' = \begin{cases} t & t \leq \text{threshold} \\ \text{threshold} & t > \text{threshold} \end{cases}$$

$$\delta' = \begin{cases} \delta & t \leq \text{threshold} \\ 0 & t > \text{threshold} \end{cases}$$

- ▶ R session: analyze the `asauro::pancreatic2` dataset, truncating the analysis to the first 6 months

Where to go from here

- ▶ regression analysis
- ▶ Cox regression
 - ▶ time-dependent covariates
 - ▶ time-dependent coefficients
 - ▶ competing risks
 - ▶ left censoring
 - ▶ multiple events
- ▶ parametric models for censored duration data

107 / 129

108 / 129

- ▶ generally speaking, if we have n observations, we can only estimate a model with *at most* $p = n$ parameters
- ▶ if we have many features p , with $p \gg n$, we can apply more general *machine learning* techniques (features selection, random forest, ...)
- ▶ we are still able though to fit Cox models using **penalized regression**

109 / 129

Elastic Net Cox model (cont.)

- ▶ a suitable value of c can be selected by e.g. cross-validation
- ▶ for $\alpha = 1$, we have the special case of the *lasso* penalty
- ▶ there is a *very fast* implementation available in the R package `glmnet`, by the same authors of the method: Jerome Friedman, Trevor Hastie and Rob Tibshirani

References: Simon, N., Friedman, J., Hastie, T., Tibshirani, R. (2011) Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent, Journal of Statistical Software, Vol. 39(5) 1-13

111 / 129

- ▶ Remember the **partial likelihood** from the Cox model:

$$l(\beta) = \prod_{j=1}^D \frac{\exp(x'_{i(j)}\beta)}{\sum_{k \in R_j} \exp(x'_k\beta)}$$

When $p > n$, the β which maximizes it goes to $+\infty$

- ▶ we thus introduce the following **elastic net** constraint on β :

$$\alpha \sum |\beta_i| + (1 - \alpha) \sum \beta_i^2 \leq c$$

for some pre-specified value of c , and some pre-set weight α

110 / 129

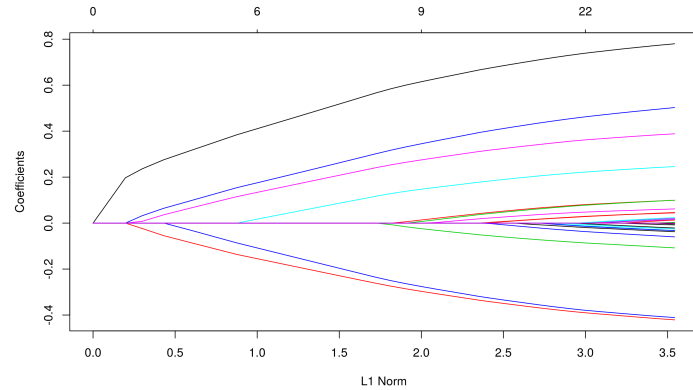
Penalized Cox regression in R

```

1 library(survival)
2 library(glmnet)
3
4 set.seed(1234)
5
6 N <- 1000 # sample size
7 p <- 30 # num. features
8 nzc <- p/3 # num. 'true' predictors
9
10 x <- matrix(rnorm(N * p), nrow = N, ncol = p)
11 beta <- rnorm(nzc)
12 linear_predictor <- x[, seq_len(nzc)] %*% beta / 3
13
14 hazard <- exp(linear_predictor)
15
16 y_time <- rexp(N, rate = hazard)
17 y_cens <- rbinom(n = N, prob = 0.3, size = 1)
18 y <- Surv(y_time, 1 - y_cens)
19
20 fit <- glmnet(x, y, family="cox")
21 plot(fit)

```

112 / 129



113 / 129

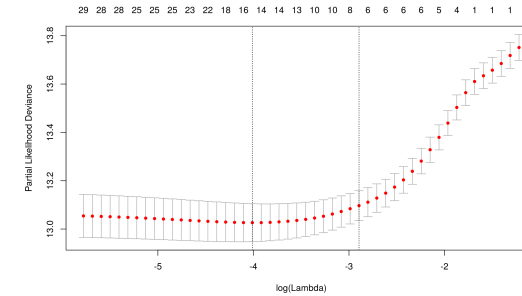
Cross validation results

```
1 str(fit.cv10)
```

```
1 List of 10
2 $ lambda      : num [1:50] 0.295 0.269 0.245 0.223 0.203 ...
3 $ cvm         : num [1:50] 13.8 13.7 13.7 13.7 13.6 ...
4 $ cvsd        : num [1:50] 0.0541 0.0533 0.0533 0.0534 0.0534 ...
5 $ cvup        : num [1:50] 13.8 13.8 13.7 13.7 13.7 ...
6 $ cvlo        : num [1:50] 13.7 13.7 13.6 13.6 13.6 ...
7 $ nzzero      : Named int [1:50] 0 1 1 1 1 1 4 4 5 5 ...
8 ..- attr(*, "names")= chr [1:50] "s0" "s1" "s2" "s3" ...
9 ...
10 $ lambda.min : num 0.0181
11 $ lambda.1se : num 0.0553
```

115 / 129

```
1 set.seed(1234)
2
3 fit.cv10 <- cv.glmnet(x, y, family = "cox")
4 plot(fit.cv10)
```



114 / 129

Predictions

```
1 coef(fit.cv10, s = "lambda.1se")
2 ##                1
3 ## V1      0.58428498
4 ## V2      .
5 ## V3      .
6 ## V4      0.31709716
7 ## V5      0.12829144
8 ## V6      0.25333876
9 ## V7      .
10 ## V8     -0.27412086
11 ## ...
13 predict(fit.cv10,
14         newx = x[1:5, ],
15         s = "lambda.1se")
16 ##                1
17 ## [1,] -1.3542387
18 ## [2,]  0.1777181
19 ## [3,]  0.7534189
20 ## [4,] -0.6364879
21 ## [5,]  0.6758198
```

```
1 b <- coef(fit.cv10, s = "lambda.1se")
2 b.i <- which(b!=0)
3 bnz <- b[b.i]
4 y0 <- x[1:5, b.i, drop = FALSE] %
5     *% bnz
```

116 / 129

- ▶ `LymphomaData.rda`
 - ▶ `x`: gene expression matrix: 7399 genes \times 240 samples
 - ▶ `time`: survival times
 - ▶ `status`: censoring status: 1 = observed, 0 = censored
- ▶ Use `glmnet` to fit a Cox model to find a predictor of survival based on gene expression
- ▶ Split the data into a training set, where you develop the model, and a testing set, where the model performance is assessed

Risk biomarkers for CRC



ARTICLE

Test of Four Colon Cancer Risk-Scores in Formalin Fixed Paraffin Embedded Microarray Gene Expression Data

Antonio F. Di Narzo, Sabine Tejpar, Simona Rossi, Pu Yan, Vlad Popovici, Pratyaksha Wirapati, Eva Budinska, Tao Xie, Heather Estrella, Adam Pavlicek, Mao Mao, Eric Martin, Weinrich Scott, Fred T. Bosman, Arnaud Roth, Mauro Delorenzi

Manuscript received December 9, 2013; revised April 22, 2014; accepted July 2, 2014.

Correspondence to: Mauro Delorenzi, PhD, SIB Swiss Institute of Bioinformatics, and University Lausanne, Office 2021, G  nopode-UNIL, Quartier Sorge, CH-1015 Lausanne, Switzerland (e-mail: mauro.delorenzi@unil.ch).

Molecular markers of risk

Table 1. Description of the four risk scores analyzed*

Abbreviation	Risk scores			
	GHS	VDS	MDA	ALM
Developer	Genomic Health	Veridex	MD Anderson	ALMAC diagnostics
Type of assay	Q-RT-PCR	microarray and Q-RT-PCR	microarray	microarray
Type of tissue	FFPE	fresh frozen and FFPE	fresh frozen	FFPE
Main publication	O'Connell et al. 2010.	Jiang et al. 2008.	Oh et al. 2011.	Kennedy et al. 2011.
Total number of features	7	7	114 (86 genes)	634 (482 genes)
Features used (genes)	7	6	85 (85 genes)	634 (identical platform)

* ALM = the scoring system proposed by Almac researchers; GHS = scoring system proposed by Genomic Health researchers; FFPE = formalin fixed paraffin embedded; MDA = scoring system proposed by researchers from MD Anderson Cancer Center; Q-RT-PCR = quantitative real-time PCR (Q-RT-PCR); VDS = scoring system proposed by Veridex researchers.

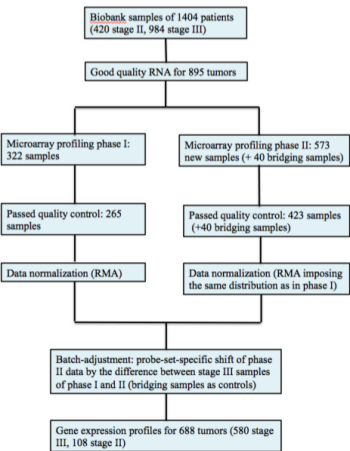


Table 2. Cox models estimates					
Outcome	Marker	Univariate*		Multivariable**	
		HR (95% CI)	P‡	HR (95% CI)	P‡
RFS	aGHS	1.33 (1.13 to 1.56)	<.001	1.30 (1.11 to 1.53)	.001
	aVDS	1.29 (1.10 to 1.52)	.002	1.27 (1.07 to 1.51)	.007
	aMDA	1.10 (0.93 to 1.30)	.26	1.13 (0.93 to 1.37)	.22
	ALM	1.31 (1.13 to 1.53)	<.001	1.20 (1.02 to 1.40)	.03
	CS4§	1.56 (1.33 to 1.84)	<.001	1.45 (1.23 to 1.71)	<.001
SAR	aGHS	1.16 (0.95 to 1.43)	.14	1.16 (0.92 to 1.46)	.20
	aVDS	0.90 (0.72 to 1.13)	.38	0.84 (0.66 to 1.08)	.17
	aMDA	1.81 (1.45 to 2.27)	<.001	1.89 (1.46 to 2.46)	<.001
	ALM	1.19 (0.97 to 1.47)	.10	1.10 (0.88 to 1.36)	.40
	CS4§	1.46 (1.18 to 1.82)	<.001	1.33 (1.05 to 1.67)	.017
OS	aGHS	1.36 (1.13 to 1.64)	.001	1.34 (1.10 to 1.62)	.003
	aVDS	1.24 (1.03 to 1.50)	.02	1.21 (0.99 to 1.48)	.07
	aMDA	1.31 (1.08 to 1.58)	.006	1.37 (1.09 to 1.71)	.007
	ALM	1.38 (1.16 to 1.65)	<.001	1.22 (1.02 to 1.47)	.03
	CS4§	1.74 (1.44 to 2.10)	<.001	1.57 (1.29 to 1.91)	<.001

* Cox proportional hazards regression models were used to estimate hazard ratios for one interquartile range variation of the continuous risk scores; no stratification was applied; adjustment by treatment was applied only in the multivariable models. aGHS = microarray-based approximation of the scoring system proposed by Genomic Health researchers; ALM = the scoring system proposed by Almac researchers; aMDA = approximation of the scoring system proposed by researchers from MD Anderson Cancer Center; aVDS = approximation of the scoring system proposed by Veridex researchers; CI = confidence interval; CS4 = the scoring system obtained by combining the four existing systems; HR = hazard ratio; OS = overall survival; RFS = relapse-free survival; SAR = survival after relapse.

† Each multivariable model included one gene expression risk score and the following variables: age, gender, TNM staging (T-stage, N-stage) (27), grade, location (right = proximal, left = distal), treatment arm, presence of lymphovascular invasion, and microsatellite instability (MSI) status.

‡ Shown are single-test P-values. The statistical significance cutoff by the Bonferroni principle (considering three tests) is at 0.05/3 = 0.0167.

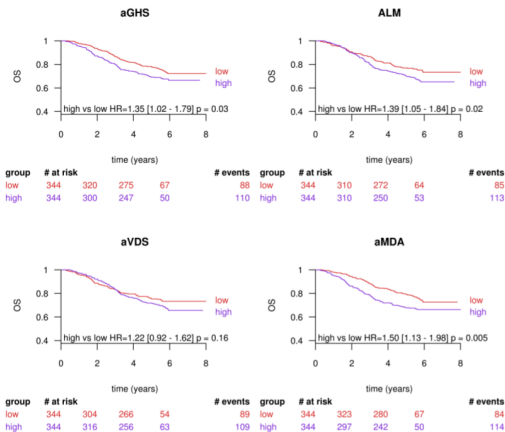


Table 3. Three-year survival				
Marker	Risk group	RFS		OS
		% (95% CI) *	% (95% CI) *	% (95% CI) *
aGHS	Whole cohort (N = 688)	66.9 (63.5 to 70.5)	34.4 (28.7 to 41.2)	83.4 (80.6 to 86.2)
	low	69.6 (64.9 to 74.7)	40.7 (32.4 to 51.2)	86.5 (83.0 to 90.2)
	high	64.2 (59.4 to 69.5)	28.5 (21.3 to 38.0)	80.2 (76.1 to 84.5)
aVDS	low	70.9 (66.2 to 75.8)	30.0 (21.8 to 41.1)	83.4 (79.5 to 87.4)
	high	63.0 (58.1 to 68.3)	37.6 (30.2 to 46.7)	83.4 (79.5 to 87.4)
aMDA	low	69.1 (64.3 to 74.1)	49.8 (41.2 to 60.1)	88.3 (84.9 to 91.8)
	high	64.8 (60.0 to 70.1)	19.9 (13.6 to 28.9)	78.5 (74.2 to 82.9)
ALM	low	70.8 (66.1 to 75.8)	36.8 (28.3 to 47.8)	86.6 (83.0 to 90.2)
	high	63.1 (58.2 to 68.4)	32.4 (25.1 to 41.6)	80.2 (76.1 to 84.5)
CS4	low	70.5 (65.8 to 75.5)	41.8 (33.1 to 52.9)	87.4 (84.0 to 91.0)
	high	63.4 (58.5 to 68.7)	28.7 (21.8 to 37.8)	79.3 (75.2 to 83.7)

* Estimated proportions of three-year survival (percentage) by the Kaplan-Meier method with 95% confidence intervals for the whole cohort and for risk groups defined by splitting the cohort at the median of each risk score into equally sized subgroups. aGHS = microarray-based approximation of the scoring system proposed by Genomic Health researchers; ALM = the scoring system proposed by Almac researchers; aMDA = approximation of the scoring system proposed by researchers from MD Anderson Cancer Center; aVDS = approximation of the scoring system proposed by Veridex researchers; CI = confidence interval; CS4 = the scoring system obtained by combining the four existing systems; OS = overall survival; RFS = relapse-free survival; SAR = survival after relapse.

Table 4. Concordance by risk score and endpoint groups*

Scoring method	Risk score subgroup	Actual survival group		
		Poor	Good	Rest
aGHS	Q1	46 (26.7%)	116 (67.4%)	10 (5.8%)
	Q2	58 (33.7%)	103 (59.9%)	11 (6.4%)
	Q3	54 (31.4%)	105 (61.0%)	13 (7.5%)
	Q4	69 (40.1%)	84 (48.8%)	19 (11.1%)
aVDS	Q1	40 (23.3%)	117 (68.0%)	15 (8.7%)
	Q2	60 (34.9%)	101 (58.7%)	11 (6.4%)
	Q3	63 (36.6%)	96 (55.8%)	13 (7.5%)
	Q4	64 (37.2%)	94 (54.7%)	14 (8.2%)
aMDA	Q1	51 (29.7%)	109 (63.4%)	12 (7.0%)
	Q2	55 (32.0%)	100 (58.1%)	17 (9.9%)
	Q3	62 (36.0%)	100 (58.1%)	10 (5.9%)
	Q4	59 (34.3%)	99 (57.6%)	14 (8.1%)
ALM	Q1	50 (29.1%)	110 (64.0%)	12 (7.0%)
	Q2	50 (29.1%)	109 (63.4%)	13 (7.6%)
	Q3	54 (31.4%)	103 (59.9%)	15 (8.7%)
	Q4	73 (42.4%)	86 (50.0%)	13 (7.5%)
CS4	Q1	36 (20.9%)	123 (71.5%)	13 (7.6%)
	Q2	65 (37.8%)	100 (58.1%)	7 (4.1%)
	Q3	57 (33.1%)	100 (58.1%)	15 (8.7%)
	Q4	69 (40.1%)	85 (49.4%)	18 (10.4%)

Table 5. Time-dependent receiver operating characteristic curves, area under curve (time = 3 years) by endpoint and risk score

Endpoint	Marker	AUC (ref. model) *	AUC gain*	P†
RFS	aGHS	0.6723	0.0136	.04
	aVDS		0.0185	.009
	aMDA		0.0085	.17
	ALM		0.0089	.16
	CS4		0.0222	.0008
SAR	aGHS	0.6406	0.0192	.11
	aVDS		0.0001	.79
	aMDA		0.0838	.0001
	ALM		0.0053	.54
	CS4		0.0443	.005
OS	aGHS	0.6918	0.0187	.005
	aVDS		0.0135	.03
	aMDA		0.0243	.001
	ALM		0.0140	.02
	CS4		0.0403	.0001

* Area under curve (AUC) for predicting survival status at three years was computed by risk scoring methods and endpoint. A reference model was fitted using the predictor variables N-stage, T-stage, and MSI status. The AUC gain was computed by adding the gene expression risk score to the predictor variables in the model. aGHS = microarray-based approximation of the scoring

Supplementary Table 3. Concordance Index gains for risk scores by endpoint.

Endpoint	marker	concordance index*		
		clinical only†	difference	p-value‡
RFS		0.6432		
	aGHS		0.0115	0.015
	aVDS		0.0154	0.001
	aMDA		0.0070	0.10
	ALM		0.0092	0.04
SAR	CS4	0.5930	0.0229	0.0001
	aGHS		0.0069	0.35
	aVDS		0.0089	0.23
	aMDA		0.0615	0.0001
	ALM		0.0022	0.69
OS	CS4	0.6620	0.0201	0.016
	aGHS		0.0144	0.003
	aVDS		0.0108	0.017
	aMDA		0.0147	0.004
	ALM		0.0095	0.03
	CS4		0.0270	0.0001

- Analyze a **right-censored** survival dataset of your choice and apply some of the methods introduced in this course: nonparametric estimation, Logrank test, Cox regression, machine learning + validation
- You can use the dataset pbc from the survival R package. See ?pbc for a detailed description of all the variables. If you want, you can pick any other dataset with **right censored** survival data
- Produce a **pdf** report with:
 - **MAX 20 PAGES**
 - Brief description of the data
 - Basic descriptive statistics (sample size, variables min/max, categorical variables distribution, etc.)
 - questions asked, methods used, results
 - please include R code either in an appendix or inline with the main report
 - you might use an Rstudio notebook, but please only send the compiled pdf report

Project (cont.)

- ▶ you can form teams of 2/3 students each
- ▶ the work can be *machine learning* oriented: i.e. build a predictor for survival; if so, show how the predictor is built, and assess its performance through survival curves, Cox Regression, etc.
- ▶ in general: analyze a survival dataset using the skills learned in this class
- ▶ due date: [see moodle](#)