STAD–Survival analysis

Mathieu Ribatet—Full Professor of Statistics

References

[1] D. Collett. *Modelling Survival Data in Medical Research*. Chapman and Hall/CR, 3rd edition, 2014.

[2] J. P. Klein and M. L. Moeschberger. *Survival analysis*. Springer–Verlag, 2nd edition, 2003.

[3] D. G. Kleinbaum and M. Klein. *Survival analysis: A self learning text*. Springer–Verlag, New–York, 3rd edition, 2012.

What is survival analysis?

- Survival analysis is the analysis of times from a time origin until the occurrence of some event or end–point.
- Typical examples are:
  - recruitement of an individuals (time origin) for an experimental study and first occurence of some symptoms
  - begining of a stress testing (time origin) up to default of an electrical component
- Essentialy you can use survival analysis as long as the data you analyzed are durations.
- It is widely used in medical research but (unfortunately) less used in quality control.

What is survival analysis?

- Survival analysis is the analysis of times from a time origin until the occurrence of some event or end–point.
- Typical examples are:
  - recruitement of an individuals (time origin) for an experimental study and first occurence of some symptoms
  - begining of a stress testing (time origin) up to default of an electrical component
- Essentialy you can use survival analysis as long as the data you analyzed are durations.
- It is widely used in medical research but (unfortunately) less used in quality control.

☞ Hence do not panick if I use medical datasets and I mainly use "medical phrasing" ;-)

Special features of survival data

- We cannot use standard statistical techniques for survival data due to
  - asymmetry, e.g., positively skewed, since durations are obviously positive[1]
  - censoring, i.e., some observation are not "real" observation (I will be more specific later)
- We thus need a new framework for these data!

[1] one could take the log of the data and analyze the transformed data but it is more sensible to work on the original scale

Examples

`genfan` Failure times of diesel engine fans
`imotor` Failure times of motor insulation
`ovarian` Survival times on two different treatments for ovarian cancer
`Rossi` Time to no recidivism when there is financial support or no financial support on releas

Examples

`genfan` Failure times of diesel engine fans
`imotor` Failure times of motor insulation
`ovarian` Survival times on two different treatments for ovarian cancer
`Rossi` Time to no recidivism when there is financial support or no financial support on releas
`ours` we will create our own data

Examples

`genfan` Failure times of diesel engine fans
`imotor` Failure times of motor insulation
`ovarian` Survival times on two different treatments for ovarian cancer
`Rossi` Time to no recidivism when there is financial support or no financial support on releas
`ours` we will create our own data in a few moment!

The genfan dataset

Data set 'genfan': Time to failure of 70 diesel engine fans.

- 'hours': hours of service
- 'status': 1=failure, 0=censored

```
> head(genfan)
  hours status
1   450      1
2   460      0
3  1150      1
4  1150      1
5  1560      0
6  1600      1
```

The imotor dataset

Data set 'imotor': Breakdown of motor insulation as a function of temperature.

- temp: temperature of the test
- time: time to failure or censoring
- status: 0=censored, 1=failed

```
> head(imotor)
  temp time status
1  150 8064      0
2  150 8064      0
3  150 8064      0
4  150 8064      0
5  150 8064      0
6  150 8064      0
```

The ovarian dataset

```
Ovarian Cancer Survival Data

Description:

     Survival in a randomised trial comparing two treatments for
     ovarian cancer

Format:

       futime:    survival or censoring time
       fustat:    censoring status
       age:       in years
       resid.ds:  residual disease present (1=no,2=yes)
       rx:        treatment group
       ecog.ps:   ECOG performance status (1 is better, see reference)

head(ovarian)
  futime fustat     age resid.ds rx ecog.ps
1     59      1 72.3315        2  1       1
2    115      1 74.4932        2  1       1
3    156      1 66.4658        2  1       2
4    421      0 53.3644        2  2       1
5    431      1 50.3397        2  1       1
6    448      0 56.4301        1  1       2
```

The `Rossi` dataset

```
Description:

     Time to recidivism on 432 convicts who were released from
     Maryland in the 70s (half of them get financial support)

Format:

       week:      time to recidivism or censoring time
       arrest:    arrtested or not, i.e., censoring status
       fin:       is there a financial support
       age:       age (years) at release
       race:      black or other
       wexp:      has a working experience after release
       mar:       is married on release
       paro:      was released on parole
       prio:      number of prior conviction
       educ:      a factor encoding the level of education

> head(data)
  week arrest fin age  race wexp         mar paro prio educ
1   20      1  no  27 black   no not married  yes    3    3
2   17      1  no  18 black   no not married  yes    8    4
3   25      1  no  19 other  yes not married  yes   13    3
4   52      0 yes  23 black  yes     married  yes    1    5
```

Let's introduce ourself

- Who am I?
- Who are you?
- What do you know about statistics and probability?

The Kaplan–Meier theatre (from Thomas A. Gerds)

- We are all on the Titanic, and the Titanic is going down. Once under water, you would have to hold your breath. But how long can you do this?
- Pick up a sheet of paper and write:
  - an unique identifier
  - your sex, smoking status, un/like sport
- Pick up your smartphone and open a timer
- The Titanic is about to sink and as I say your personal id you should start holding your breath
- If you stop holding your breath at some time, write it down on your sheet of paper
- If you hold your breath until I said stop, just write the current time on your timer adding a plus after it, e.g., 52.3^{+}.

Let's talk about our data

1. Preliminaries

Patient time and study time

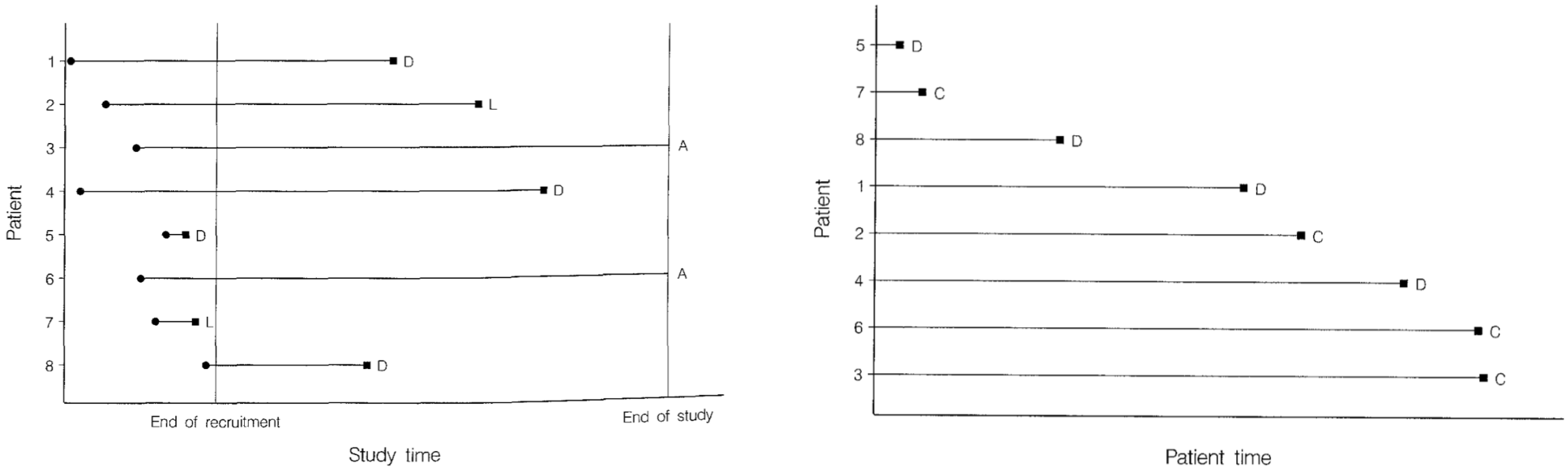


Figure 1: *Illustration of patient and study times. Left: Study time. (D: death, L: lost to follow-up, A: alive). Right: Patient time (C: right censored).*

- Often patients are not all recruited at the same time
- We have to pay attention to the difference between study and patient time

Patient time and study time

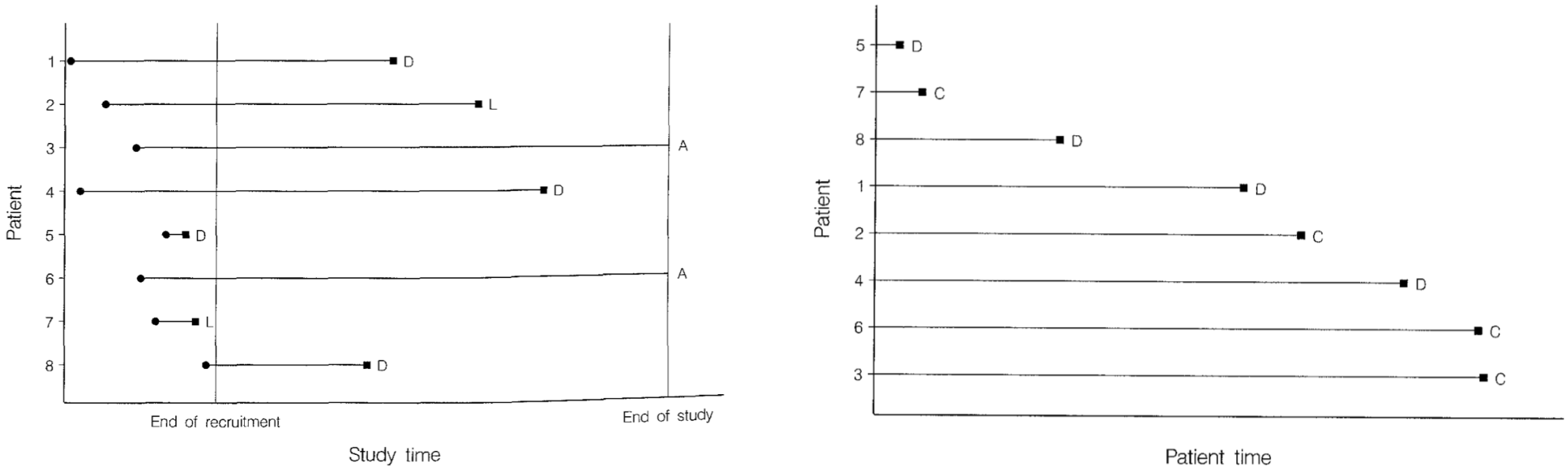


Figure 1: *Illustration of patient and study times. Left: Study time. (D: death, L: lost to follow-up, A: alive). Right: Patient time (C: right censored).*

- Often patients are not all recruited at the same time
- We have to pay attention to the difference between study and patient time
- Statistical modelling is often made on patient time

p.d.f., c.d.f. and co

- Let T be a continuous positive random variable having a probability density function f, i.e.,

$$f(t) \geq 0, \qquad \int_0^\infty f(t)\mathrm{d}t = 1.$$

- The cumulative distribution function of T is

$$F(t) = \Pr(T \leq t) = \Pr(T < t) = \int_0^t f(u)\mathrm{d}u,$$

and gives the probability that the survival time is less than t.

- The survival function of T is

$$S(t) = \Pr(T > t) = \int_t^\infty f(u)\mathrm{d}u = 1 - F(t),$$

and gives the probability that the survival time is greater than t.

Hazard rate and cumulative hazard

- In survival analysis we often talk about the hazard rate

$$h(t) = \lim_{\Delta t \downarrow 0} \frac{\Pr(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t},$$

and the cumulative hazard

$$H(t) = \int_0^t h(u) \mathrm{d}u.$$

Exercise 1. If T has density f, show that we have for all $t > 0$

$$h(t) = \frac{f(t)}{S(t)}, \qquad S(t) = \exp\{-H(t)\}.$$

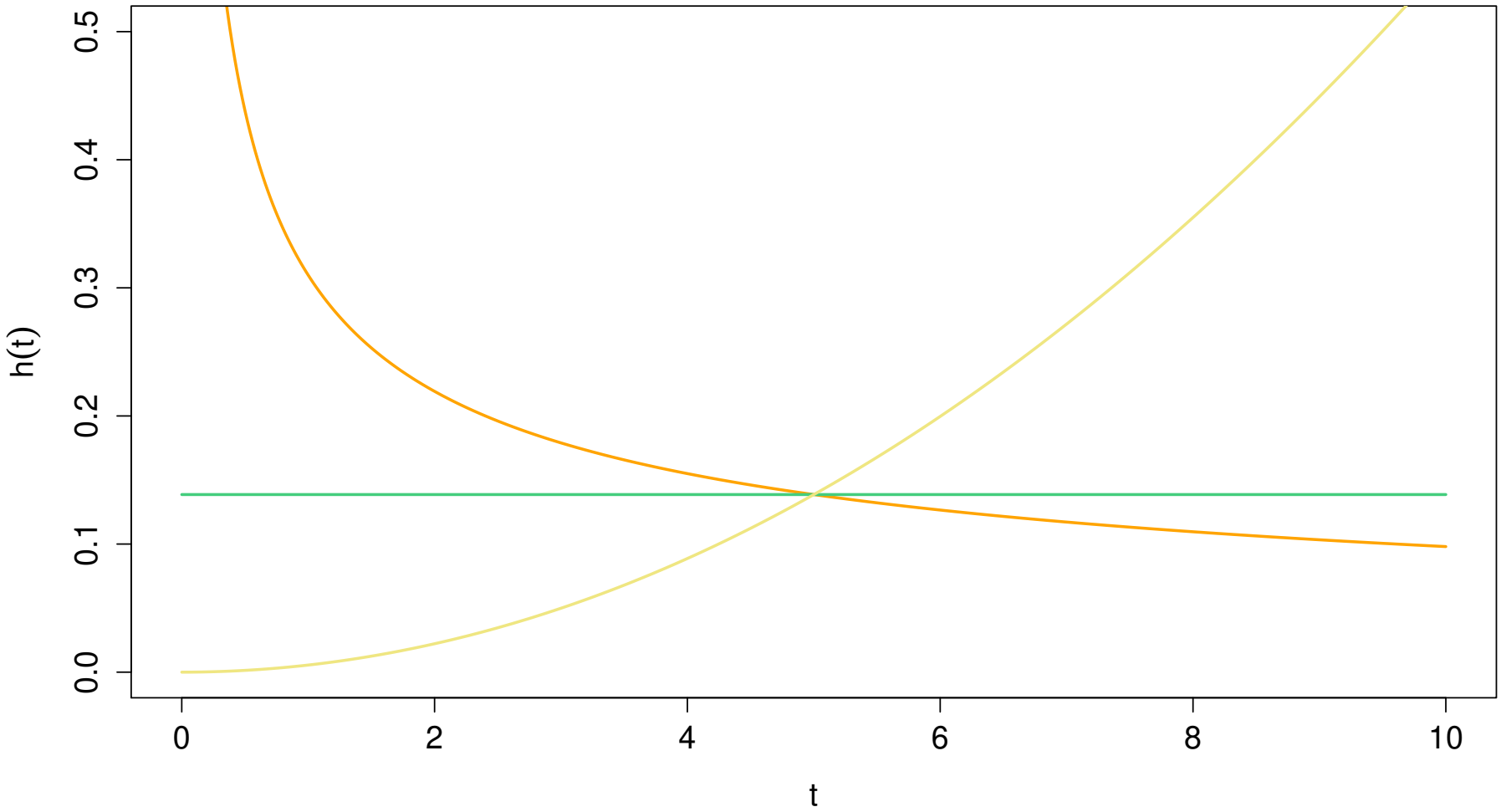


Figure 2: *Plot of three different hazard rate (obtained from a Weibull distribution).*

Censoring

Definition 1. An observation is said to be censored when the end–point has not been observed.

Example 1. lost to follow–up, child knows knows how to read before school, component still OK as the stress testing phase ends, end–point is not related to our analysis...

- There are 3 different type of censoring:
  - right censoring where we observed $T = \min(C, T_*)$;
  - left censoring where we observed $T = \max(C, T_*)$;
  - and interval censoring where we just know $T_* \in [A, B]$.

Censoring

Definition 1. An observation is said to be censored when the end–point has not been observed.

Example 1. lost to follow–up, child knows knows how to read before school, component still OK as the stress testing phase ends, end–point is not related to our analysis...

- There are 3 different type of censoring:
  - right censoring where we observed $T = \min(C, T_*)$;
  - left censoring where we observed $T = \max(C, T_*)$;
  - and interval censoring where we just know $T_* \in [A, B]$.

☞ The most common censoring is by far right censoring.

2. Non parametric estimation

Non parametric statistics

- Most often stochastic modelling assumes that the random variable T under study has a pre–specified distribution, e.g., $T \sim$ Log–Normal.
- Non parametric statistics make no distributional assumptions on T
- The price to pay is most often the speed of convergence, statistical power, e.g., larger variance of the estimator...
- However non parametric procedures are appealing as they can be used later for model checking.

Naive estimate

- Remember the empirical distribution function

$$\hat{F}(t) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{T_i \leq t\}}, \qquad t > 0,$$

where $T_i \overset{\text{iid}}{\sim} F$.

- The law of large number implies, provided $\mathbb{E}(T) < \infty$,

$$\hat{F}(t) \longrightarrow \mathbb{E}\left[1_{\{T \leq t\}}\right] = F(t), \qquad n \to \infty, \quad t > 0.$$

- Equivalently we can estimate the survivor function from

$$\hat{S}(t) = 1 - \hat{F}(t) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{T_i > t\}}.$$

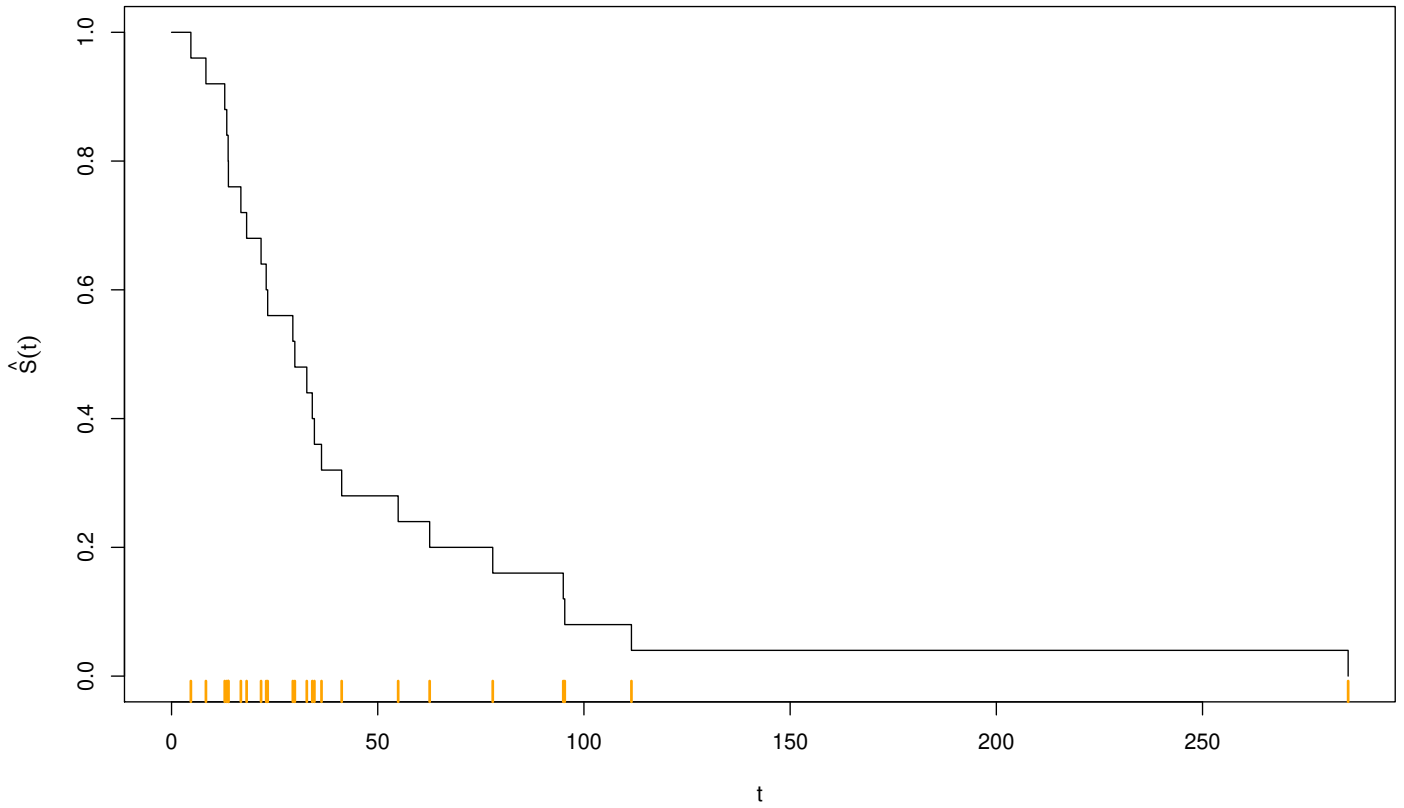


Figure 3: *An example of a naive estimate of the survival function.*

- It is a step function (actually cadlag).
- We always have $\hat{S}(0) = 1$ and $\hat{S}(\max T_i) = 0$.

There's a trap!

- The above estimator is usually consistent and unbiased.
- But it is not the case anymore when there are censored observations!
- Why? Essentially because some $T_i \not\sim F \ldots$
- We need to use different estimators for $S(t)$:
  - Kaplan–Meier estimator
  - Nelson–Aalen estimator
  - Life–table estimator
- We will cover these estimators in turn.

Towards the Kaplan–Meier estimator

- Consider the order statistics (without ties) of n durations (censored or not) $t_1, \ldots, t_n$, i.e., $0 = t_{(0)} < t_{(1)} < \cdots < t_{(n)}$.
- We thus have for any $j \in \{1, \ldots, n\}$

$$\begin{aligned}
\Pr(T > t_{(j)}) &= \Pr(T > t_{(j)}, T > t_{(j-1)}) \\
&= \Pr(T > t_{(j)} \mid T > t_{(j-1)}) \Pr(T > t_{(j-1)}) \\
&= \cdots \\
&= \prod_{\ell=1}^{j} \Pr(T > t_{(\ell)} \mid T > t_{(\ell-1)}).
\end{aligned}$$

- The Kaplan–Meier estimator uses this decomposition and plugs in empirical estimator for these conditional probabilities

Kaplan–Meier estimator

- Each conditional probability $p_\ell = \Pr(T > t_{(\ell)} \mid T > t_{(\ell-1)})$ is easily estimated from its empirical counterpart

$$\hat{p}_\ell = 1 - \frac{d_\ell}{n_\ell}, \quad d_\ell = \#\text{ deaths at } t_{(\ell)}, \quad n_\ell = \#\text{ at risk at } t_{(\ell)}^-.$$

- This gives the Kaplan–Meier estimator

$$\hat{S}_{KM}(t) = \prod_{i:\, T_i \leq t} \left(1 - \frac{d_i}{n_i}\right), \qquad t > 0.$$

Application: failure times (hours) of diesel engine fans

450	1850+	2030+	3000+	4150+	4300+	5000+	6450+	8100+	8750+
460+	1850+	2070	3000+	4150+	4600	5000+	6450+	8200+	9400+
1150	1850+	2070	3100	4150+	4850+	6100+	6700+	8500+	9900+
1150	1850+	2080	3200+	4150+	4850+	6100	7450+	8500+	10100+
1560+	1850+	2200+	3450	4300+	4850+	6100+	7800+	8500+	10100+
1600	2030+	3000+	3750+	4300+	4850+	6100+	7800+	8750+	10100+
1660+	2030+	3000+	3750+	4300+	5000+	6300+	8100+	8750	11500+

Application: failure times (hours) of diesel engine fans

450	1850+	2030+	3000+	4150+	4300+	5000+	6450+	8100+	8750+
460+	1850+	2070	3000+	4150+	4600	5000+	6450+	8200+	9400+
1150	1850+	2070	3100	4150+	4850+	6100+	6700+	8500+	9900+
1150	1850+	2080	3200+	4150+	4850+	6100	7450+	8500+	10100+
1560+	1850+	2200+	3450	4300+	4850+	6100+	7800+	8500+	10100+
1600	2030+	3000+	3750+	4300+	4850+	6100+	7800+	8750+	10100+
1660+	2030+	3000+	3750+	4300+	5000+	6300+	8100+	8750	11500+

Table 1: *Kaplan–Meier estimate of the survival function on the* `genfan` *dataset.*

Time	n_{risk}	n_{event}	$\hat{S}_{KM}$
0	70	0	1.00
450	70	1	0.99
1150	68	2	0.96
1600	65	1	0.94
2070	55	2	0.91
2080	53	1	0.89
3100	47	1	0.87
3450	45	1	0.85
4600	34	1	0.83
6100	26	1	0.80
8750	9	1	0.71

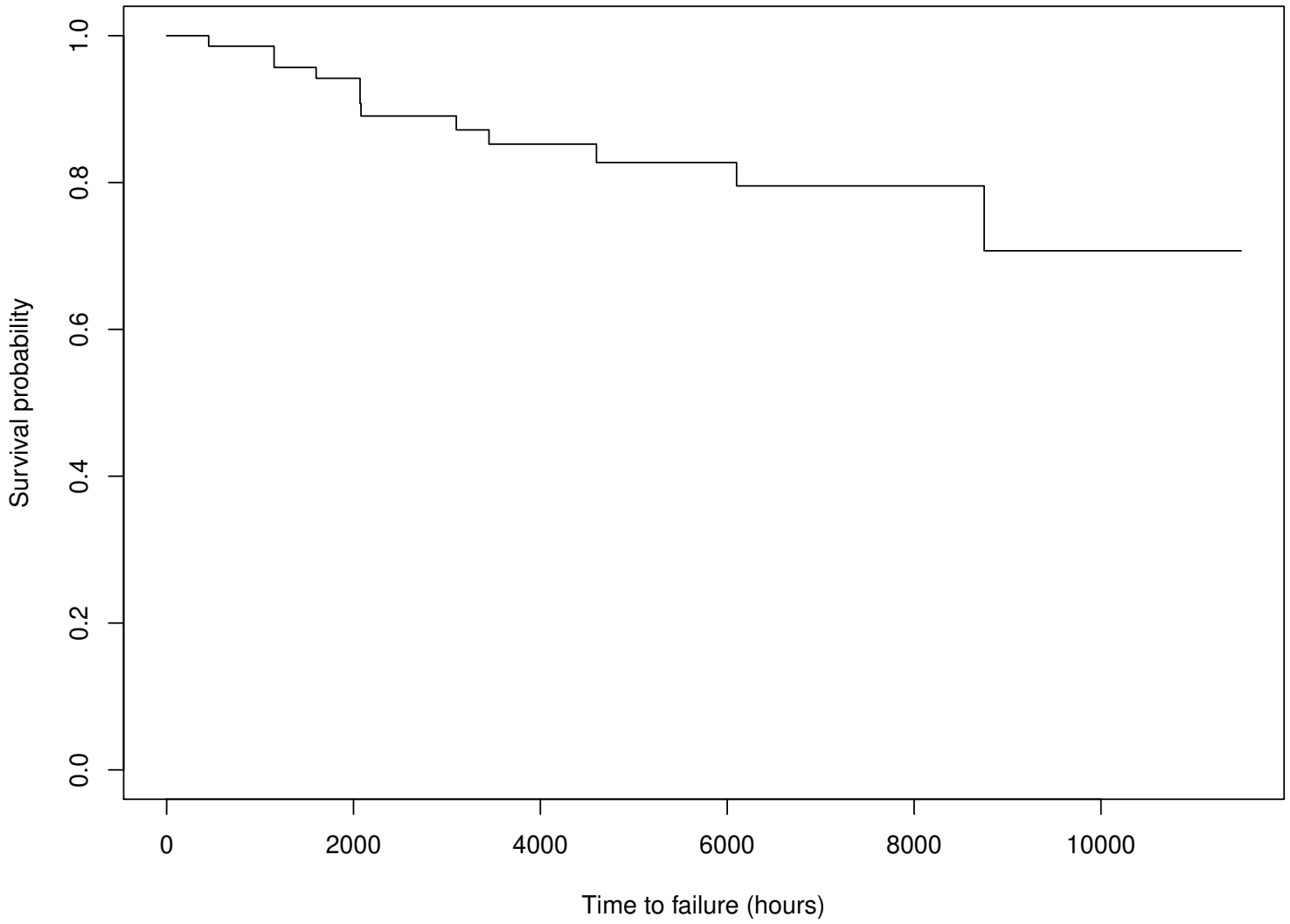


Figure 4: *Kaplan–Meier estimate for the* `genfan` *dataset.*

Ties in Kaplan–Meier

- We may have ties, e.g, $t_1 = t_2$, and in such situations we have
  - $d_1 = 2$ if obs. 1 and 2 are "death"
  - if obs. 1 is censored and 2 isn't, we assume that "death" happens before

Table 2: *Kaplan–Meier estimate of the survival function on the* `genfan` *dataset.*

Time	n_{risk}	n_{event}	$\hat{S}_{KM}$
0	70	0	1.00
450	70	1	0.99
1150	68	2	0.96
1600	65	1	0.94
2070	55	2	0.91
2080	53	1	0.89
3100	47	1	0.87
3450	45	1	0.85
4600	34	1	0.83
6100	26	1	0.80
8750	9	1	0.71

Theorem 1. *Let $X_1, X_2, \ldots$ a sequence of random variables such that*

$$\sqrt{n}(X_n - \theta) \xrightarrow{d.} N(0, \sigma^2), \qquad n \to \infty,$$

for some fixed values $\theta \in \mathbb{R}$. Let g be a function such that $g'(\theta)$ exists and is non null. Then

$$\sqrt{n}(g(X_n) - g(\theta)) \xrightarrow{d.} N(0, g'(\theta)^2\sigma^2), \qquad n \to \infty.$$

Proof. Taylor expansion + Continuous mapping theorem + Slutsky's lemma □

Delta–method

Theorem 1. *Let $X_1, X_2, \ldots$ a sequence of random variables such that*

$$\sqrt{n}(X_n - \theta) \xrightarrow{d.} N(0, \sigma^2), \qquad n \to \infty,$$

for some fixed values $\theta \in \mathbb{R}$. Let g be a function such that $g'(\theta)$ exists and is non null. Then

$$\sqrt{n}(g(X_n) - g(\theta)) \xrightarrow{d.} N(0, g'(\theta)^2\sigma^2), \qquad n \to \infty.$$

Proof. Taylor expansion + Continuous mapping theorem + Slutsky's lemma □

☞ The delta–method is very useful when one have to compute (asymptotic) variance of a transformation of an estimator

Simple application

Exercise 2. Let $X_1, X_2, \ldots$ be independent copies of a random variable X such that $\mu = \mathbb{E}(\log X)$ and $\sigma^2 = \mathrm{Var}(\log X) < \infty$. How would you get an approximation, i.e., as long as n is large enough, for the variance of the geometric mean of those X_i's?

Variance of Kaplan–Meier

$$\hat{S}_{KM}(t) = \prod_{i:\, T_i \leq t} \left(1 - \frac{d_i}{n_i}\right), \qquad t > 0.$$

- Computing the variance of a product is a mess but that of a sum of i.i.d. random variable is easy.

Variance of Kaplan–Meier

$$\hat{S}_{KM}(t) = \prod_{i:\, T_i \leq t} \left(1 - \frac{d_i}{n_i}\right), \qquad t > 0.$$

- Computing the variance of a product is a mess but that of a sum of i.i.d. random variable is easy. Idea: take the $\log$ and use Delta–method!
- Doing so we get the Greenwood formula

$$\widehat{\mathsf{Var}}(\hat{S}_{KM}(t)) = \hat{S}_{KM}(t)^2 \sum_{i:\, T_i \leq t} \frac{d_i}{n_i(n_i - d_i)}.$$

- Since Kaplan–Meier is asymptotically normal, a pointwise confidence interval for $S(t)$ of level α is

$$\left[\hat{S}_{KM}(t) - z_{1-(1-\alpha)/2}\sqrt{\widehat{\mathsf{Var}}(\hat{S}_{KM}(t))}, \hat{S}_{KM}(t) + z_{1-(1-\alpha)/2}\sqrt{\widehat{\mathsf{Var}}(\hat{S}_{KM}(t))}\right],$$

e.g., $z_{1-(1-\alpha)/2} = 1.96$ when $\alpha = 95\%$.

Application: failure times (hours) of diesel engine fans

Table 3: *Kaplan–Meier estimate of the survival function on the* `genfan` *dataset with confidence intervals.*

Time	n_{risk}	n_{event}	$\hat{S}_{KM}$	Std. err.	lower 95% CI	upper 95% CI
450	70	1	0.99	0.01	0.96	1.00
1150	68	2	0.96	0.02	0.91	1.00
1600	65	1	0.94	0.03	0.89	1.00
2070	55	2	0.91	0.04	0.84	0.98
2080	53	1	0.89	0.04	0.82	0.97
3100	47	1	0.87	0.04	0.79	0.96
3450	45	1	0.85	0.05	0.77	0.95
4600	34	1	0.83	0.05	0.73	0.93
6100	26	1	0.80	0.06	0.69	0.92
8750	9	1	0.71	0.10	0.54	0.93

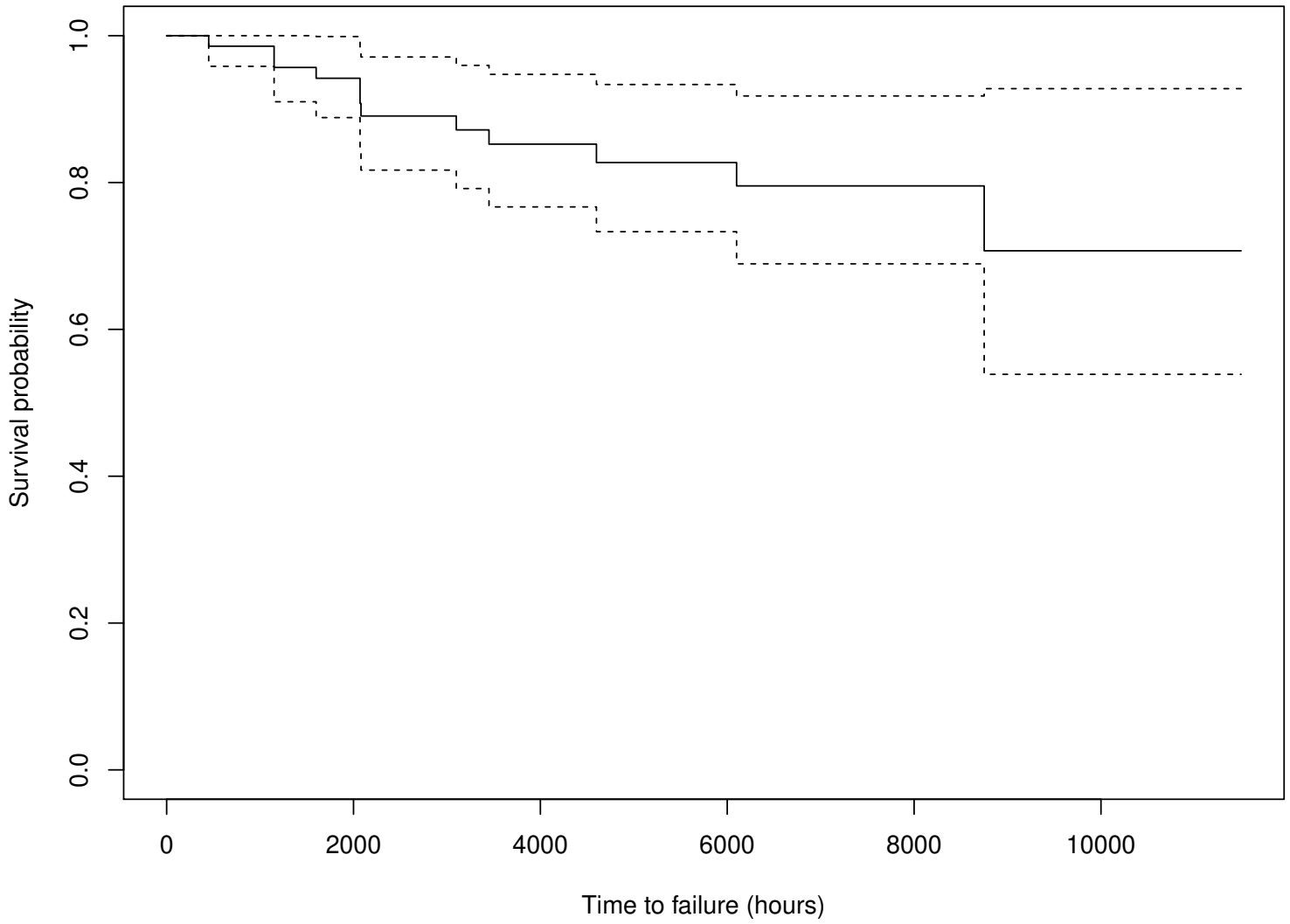


Figure 5: *Kaplan–Meier estimate for the* `genfan` *dataset.*

Derivation (quite) of the Greenwood's formula

$$\hat{S}_{KM}(t) = \prod_{i:\, T_i \leq t} \left(1 - \frac{d_i}{n_i}\right) = \prod_{i:\, T_i \leq t} \hat{p}_i \Longleftrightarrow \log \hat{S}_{KM}(t) = \sum_{i:\, T_i \leq t} \log \hat{p}_i,$$

where $\hat{p}_i = (n_i - d_i)/n_i$ is an estimator of p_i the "true" probability to survive in the i–th interval. So...

Wait a minute...

$$\left[\hat{S}_{KM}(t) - \Delta_\alpha, \hat{S}_{KM}(t) + \Delta_\alpha\right]$$

- One way to fix this is to set the lower bound to $\max\{0, \hat{S}_{KM}(t) - \Delta_\alpha\}$ and the upper bound to $\min\{1, \hat{S}_{KM}(t) + \Delta_\alpha\}$.
- Another strategy is to use the log-log survivorship function, i.e.,

$$\hat{S}_{KM}(t) \longmapsto \log\{-\log \hat{S}_{KM}(t)\},$$

  which is a one–one mapping that maps $[0, 1]$ to $\mathbb{R}$.
- Again using our best friend Delta–method, we will have a confidence interval that always belongs to $[0, 1]$.

- The Greenwood formula uses the following estimator

$$\widehat{\mathsf{Var}}(\log \hat{S}_{KM}(t)) = \sum_{i:\, T_i \leq t} \frac{d_i}{n_i(n_i - d_i)}$$

- This suggests following estimator for $\log\{-\log S(t)\}$

$$\Delta^2 := \widehat{\mathsf{Var}}(\log\{-\log \hat{S}_{KM}(t)\}) = \left(-\frac{1}{\log \hat{S}_{KM}(t)}\right)^2 \widehat{\mathsf{Var}}(\log \hat{S}_{KM}(t)).$$

- A symmetric confidence interval for $\log\{-\log S(t)\}$ is thus

$$\left[\log\{-\log \hat{S}_{KM}(t)\} - z_{1-(1-\alpha)/2}\Delta, \log\{-\log \hat{S}_{KM}(t)\} + z_{1-(1-\alpha)/2}\Delta\right],$$

- Recall that

$$IC = \left[\log\{-\log \hat{S}_{KM}(t)\} - z_{1-(1-\alpha)/2}\Delta, \log\{-\log \hat{S}_{KM}(t)\} + z_{1-(1-\alpha)/2}\Delta\right],$$

which by construction satisfies

$$\Pr(\log\{-\log S(t)\} \in IC) \longrightarrow \alpha, \qquad n \to \infty.$$

- Now switching back to the original scale we get, i.e., applying pointwise $f : x \mapsto \exp\{-\exp(x)\}$, we get

$$\Pr\{S(t) \in f(IC)\} \longrightarrow \alpha, \qquad n \to \infty,$$

where

$$f(IC) = \left[\hat{S}_{KM}(t)^{\exp(z_{1-(1-\alpha)/2}\Delta)}, \hat{S}_{KM}(t)^{\exp(-z_{1-(1-\alpha)/2}\Delta)}\right]$$

Some key features of Kaplan–Meier

- If we have no censoring, Kaplan–Meier corresponds to the usual empirical survivor function
- Let $T_{(n)}$ be the largest observation. There are two cases:
  - it is an event, then the last term is obviously

$$1 - \frac{1}{1} = 0,$$

    and $\hat{S}_{KM}(t) = 0$ for all $t \geq T_{(n)}$
  - it is a censored observation, then the last term will be stricly positive and $\hat{S}_{KM}(t) > 0$ for all $t \geq T_{(n-1)}$[2]

[2] where we suppose that $T_{(n-1)}$ is an event otherwise move to the last event.

Life–table estimator

- The life–table esimator, a.k.a. the actuarial etimator, consists in discretizing the time domain $[0, D]$ rather than using the observed time events.
- Let $D = \cup_{j=1}^{m} I_j$, $I_j = [t'_{j-1}, t'_j)$, where $t_0 = 0 < t_1 < \cdots < t_{m-1} < t_m = \infty$, and denote

$$n_j = \#\text{ alive at } t'_j, \quad d_j = \#\text{ deaths in } I_j, \quad c_j = \#\text{ censoring in } I_j.$$

- The life–table estimator assumes that within I_j the censoring process is uniform over I_j so that the expected number of observation at risk within I_j is $\tilde{n}_j = n_j - c_j/2$.
- This leads to the life–table estimator

$$\hat{S}_{LF}(t) = \prod_{i:\, T_i \leq t} \left(1 - \frac{d_i}{\tilde{n}_i}\right), \quad \widehat{\text{Var}}\{\hat{S}_{LF}(t)\} = \hat{S}_{LF}(t) \sum_{i:\, T_i \leq t} \frac{d_i}{\tilde{n}_i(\tilde{n}_i - d_i)}.$$

Nelson–Aalen estimator

- The Nelson–Aalen estimator is an estimator of the cumulative hazard, i.e.,

$$H(t) = \int_0^t \lambda(u)\mathrm{d}u, \qquad t \geq 0.$$

- The mathematical foundations for this estimator is quite technical (based on point processes and martingales) and beyond the scope of this lecture.
- The Nelson–Aalen estimator is given by

$$\hat{H}_{NA}(t) = \sum_{i:\, T_i \leq t} \frac{d_i}{n_i}, \qquad t \geq 0,$$

and an estimator for its variance is

$$\widehat{\mathsf{Var}}\{\hat{H}_{NA}(t)\} = \sum_{i:\, T_i \leq t} \frac{d_i}{n_i^2}.$$

New from old

- The Nelson–Aalen estimator can be used to estimate the survivor function since $S(t) = \exp\{-H(t)\}$, $t \geq 0$.
- This leads to the Flemming–Harington estimator

$$\hat{S}_{FH}(t) = \exp\{-\hat{H}_{NA}(t)\}, \qquad t \geq 0.$$

- Conversely one can get an estimator of the cumulative hazard using the Kaplan–Meier and life–table estimators

$$\hat{H}_{KM}(t) = -\log \hat{S}_{KM}(t), \qquad \hat{H}(t) = -\log \hat{S}_{LT}(t), \qquad t \geq 0.$$

Exercise 3. Try to give an estimate of the variance of the Flemming–Harington estimator.

As an aside

- First note that

$$\hat{S}_{FH}(t) = \exp\{-\hat{H}_{NA}(t)\} = \exp\left(-\sum_{i:\, T_i \leq t} \frac{d_i}{n_i}\right) = \prod_{i:\, T_i \leq t} \exp\left(-\frac{d_i}{n_i}\right).$$

- Since $\exp(-x) = 1 - x + o(x)$, we thus have

$$\hat{S}_{FH}(t) \approx \prod_{i:\, T_i \leq t} \left(1 - \frac{d_i}{n_i}\right) := \hat{S}_{KM}(t),$$

  as long as $d_i \ll n_i$ (which will be true except for the largest survival times).
- We can go even further by recalling that $\exp(x) \geq 1 - x$ for all $x \in \mathbb{R}$. Hence we always have $\hat{S}_{FH}(t) \geq \hat{S}_{KM}(t)$ for all $t \geq 0$.

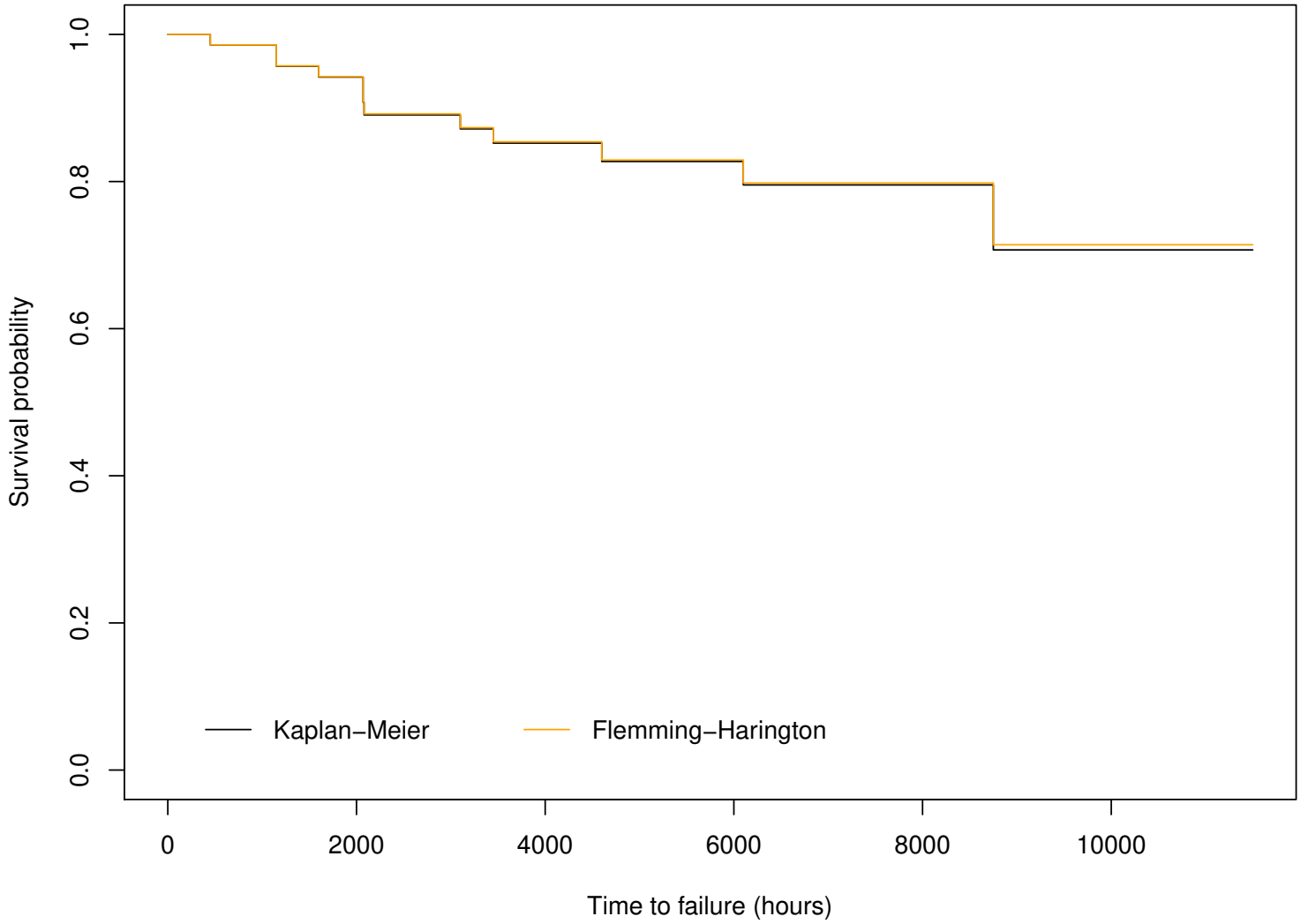


Figure 6: *Comparison of the Kaplan–Meier and Flemming–Harington estimates for the* `genfan` *dataset. As expected* $\hat{S}_{FH}(t) \geq \hat{S}_{KM}(t)$ *and both estimates are close as* $d_i \ll n_i$.

Estimating the median of survival times

- Since the distribution of survival times is often positively skewed, it is preferable to use the median rather than the expectation as a measure of location.
- The median is easily estimating from $\hat{S}(t)$ (whatever estimator you chose).
- By definition, the median, denoted by $t_{0.5}$, satisfies $S(t_{0.5}) = 0.5$ and we just substitute S for $\hat{S}$.
- Some care is needed though, since $\hat{S}$ is a step function, and we use

$$\hat{t}_{0.50} = \min\{t_i : \hat{S}(t_i) < 0.5\}.$$

- In the very unlucky situtation where $\hat{S}(t) = 0.5$ for $t \in [t_{(i)}, t_{(i+1)})$, the convention is to set

$$\hat{t}_{0.50} = \frac{t_{(i)} + t_{(i+1)}}{2}.$$

Estimating quantiles of survival times

- The above methodology easily extends to quantiles of arbitrary order p, $0 < p < 1$.
- By definition, the quantile of order p, denoted by t_p, satisfies $S(t_p) = 1 - p$.
- Hence an estimator for t_p is

$$\hat{t}_p = \min\{t_i : \hat{S}(t_i) < 1 - p\}.$$

- It may happen that $\hat{S}(t) > 1 - p$ for all $t \geq 0$. In such situtation, $\hat{t}_p$ is left undefined.

Variance of $\hat{t}_p$

- To get an estimator for $\mathrm{Var}(\hat{t}_p)$, we start by noticing that

$$\mathrm{Var}\{S(\hat{t}_p)\} \approx (-f(t_p))^2 \mathrm{Var}(\hat{t}_p).$$

- Hence we can derive an estimator for $\mathrm{Var}(\hat{t}_p)$

$$\widehat{\mathrm{Var}}(\hat{t}_p) = \frac{1}{\hat{f}(\hat{t}_p)^2} \widehat{\mathrm{Var}}\{\hat{S}(\hat{t}_p)\}.$$

- In the above estimator, $\widehat{\mathrm{Var}}\{\hat{S}(\hat{t}_p)\}$ is obtained from the Kaplan–Meier or Flemming–Harrington variance, while $\hat{f}(\hat{t}_p)$ is typically estimated using finite differences on $\hat{S}$.
- Confidence intervals are done using the usual way, i.e.,

$$\left[\hat{t}_p - z_{1-(1-\alpha)/2}\sqrt{\widehat{\mathrm{Var}}(\hat{t}_p)}, \hat{t}_p + z_{1-(1-\alpha)/2}\sqrt{\widehat{\mathrm{Var}}(\hat{t}_p)}\right]$$

Comparison of two groups of survival data

- Often one may wonder if condition A yields larger survival times than condition B.
- One (non rigourous) way is to compare the two estimated survival function $\hat{S}_A$ and $\hat{S}_B$.
- The formal way to do it is within the (statistical) hypothesis testing framework

Hypothesis testing (reminder or not)

- A statistical hypothesis testing consists of 3 ingredients:
  - Hypothesis: the null hypothesis H_0 and its alternative H_1 (usually the complement of H_0).
  - A test statistics T whose distribution is known under the null H_0
  - A binary decision rule indicating if we
    - are not able to reject H_0 in favor of H_1
    - reject H_0 in favor of H_1.

- The decision rule is always defined from a Type I error $\alpha \in (0, 1)$, e.g., $\alpha = 5\%, 10\%$ which corresponds to

$$\alpha = \Pr(\text{decide in favor of } H_1 \mid H_0 \text{ is true})$$

- The type I error α is set by the user while the Type II error β is unknown

$$\beta = \Pr(\text{not able to reject } H_0 \mid H_1 \text{ is true})$$

Plot of the decision rule

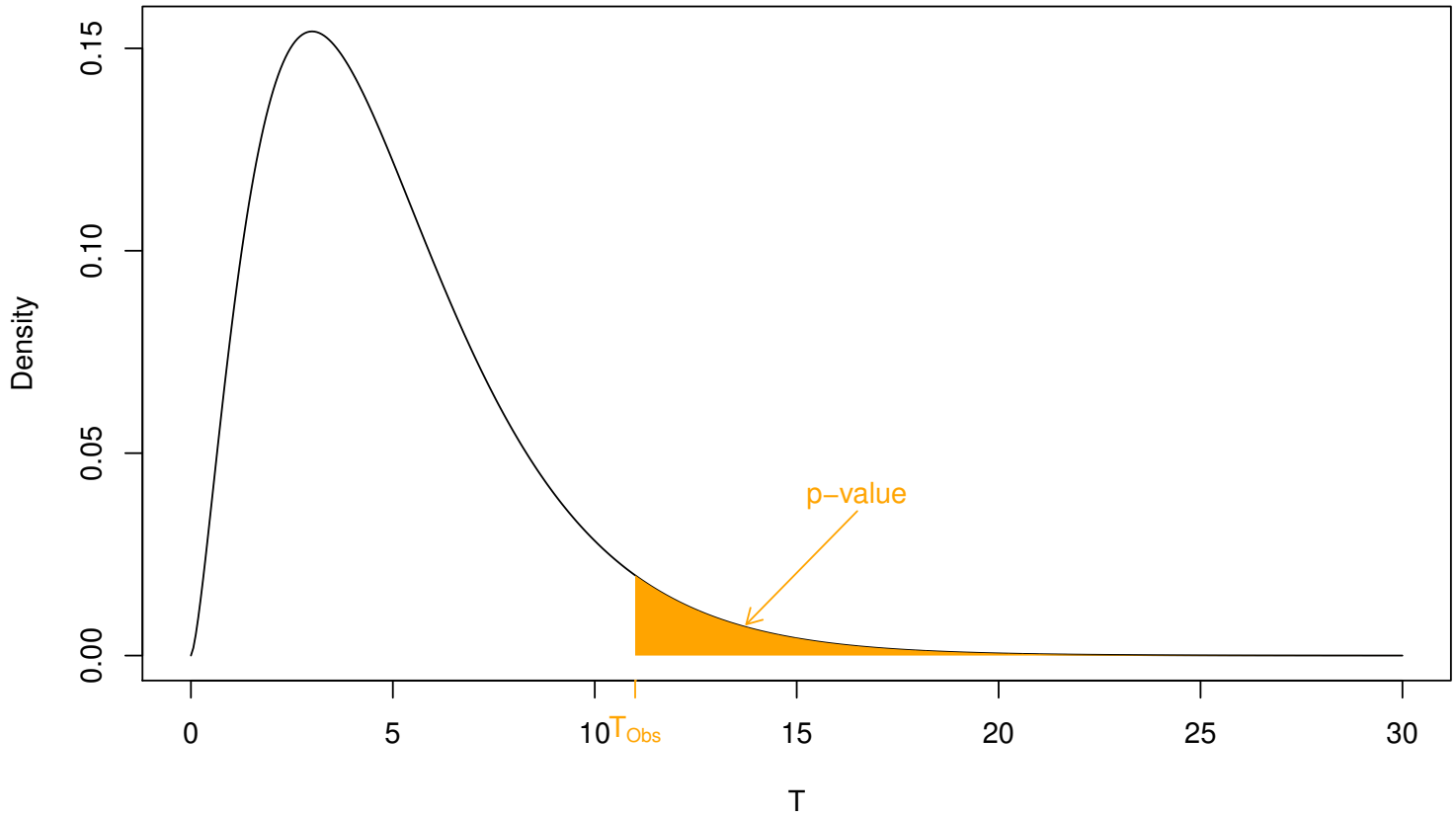


Figure 7: *Illustration of the decision rule for a hypothesis test.*

Towards the log–rank test

- Suppose we have two conditions A and B and that we focus on a specific time event $t_{(i)}$. We summarize the data as follow

Group	Number of death at $t_{(i)}$	Number of surviving beyond $t_{(i)}$	Number at risk just before $t_{(i)}$
A	d_{Ai}	$n_{Ai} - d_{Ai}$	n_{Ai}
B	d_{Bi}	$n_{Bi} - d_{Bi}$	n_{Bi}
Total	d_i	$n_i - d_i$	n_i

- The hypothesis are $H_0 : S_A = S_B$ versus $H_1 : S_A \neq S_B$.
- Now under H_0, i.e., independence on the conditions A or B,[3] we have

$$d_{Ai} \sim \text{HyperGeometric}(n_i, d_i, n_{Ai}),$$

where here n_i is the population size, d_i is the number of successes within the population and n_{Ai} the number of trials (without replacement).

[3] treating the margins of the table as fixed

The hypergeometric distribution

- The hypergeometric distribution is a discrete distribution that models the number of sucesses in successive trials without replacement.
- To be more specific, the setting is the following:
  - A finite population of size N
  - Among those N units, K have a desired feature (success if drawn)
  - The number of trials n.
- We thus write $X \sim \text{HyperGeom}(N, K, n)$ and we have

$$\Pr(X = k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}, \qquad k \in \{0, \ldots, K\},$$

which makes sense since

$$\binom{N}{n} = \qquad \binom{K}{k} = \qquad \binom{N-K}{n-k} =$$

Towards the log–rank test (2)

- If $d_{Ai} \sim \text{HyperGeometric}(n_i, d_i, n_{Ai})$, we have

$$\mu_{Ai} := \mathbb{E}(d_{Ai}) = \frac{n_{Ai} d_i}{n_i}, \qquad \sigma_i^2 := \text{Var}(d_{Ai}) = \frac{n_{Ai} n_{Bi} d_i (n_i - d_i)}{n_i^2 (n_i - 1)}.$$

- Now it makes sense to sum across all time events, leading to

$$U_L = \sum_{i=1}^{r} (d_{Ai} - \mu_{Ai})$$

Towards the log–rank test (2)

- If $d_{Ai} \sim \text{HyperGeometric}(n_i, d_i, n_{Ai})$, we have

$$\mu_{Ai} := \mathbb{E}(d_{Ai}) = \frac{n_{Ai} d_i}{n_i}, \qquad \sigma_i^2 := \text{Var}(d_{Ai}) = \frac{n_{Ai} n_{Bi} d_i (n_i - d_i)}{n_i^2 (n_i - 1)}.$$

- Now it makes sense to sum across all time events, leading to

$$U_L = \sum_{i=1}^{r} (d_{Ai} - \mu_{Ai}) = \sum_{i}^{r} (\text{Observed} - \text{Expected}),$$

where r is the number of event.

Towards the log–rank test (2)

- If $d_{Ai} \sim \text{HyperGeometric}(n_i, d_i, n_{Ai})$, we have

$$\mu_{Ai} := \mathbb{E}(d_{Ai}) = \frac{n_{Ai} d_i}{n_i}, \qquad \sigma_i^2 := \text{Var}(d_{Ai}) = \frac{n_{Ai} n_{Bi} d_i (n_i - d_i)}{n_i^2 (n_i - 1)}.$$

- Now it makes sense to sum across all time events, leading to

$$U_L = \sum_{i=1}^{r} (d_{Ai} - \mu_{Ai}) = \sum_{i}^{r} (\text{Observed} - \text{Expected}),$$

  where r is the number of event.
- It can be shown that U_L is approximately normal with mean 0, hence

$$T = \frac{U_L}{\sqrt{\sum_{i=1}^{r} \sigma_i^2}} \dot{\sim} N(0,1) \quad \text{or equivalently} \quad T^2 = \frac{U_L^2}{\sum_{i=1}^{r} \sigma_i^2} \dot{\sim} \chi_1^2.$$

The log–rank test

- The log–rank test set the hypothesis to

$$H_0 : t \mapsto S_A(t) - S_B(t) \equiv 0 \quad \text{vs.} \quad H_1 : t \mapsto S_A(t) - S_B(t) \not\equiv 0$$

- It consists in the following steps:
  - Fix the Type I error to some level α;
  - Compute the observed test statistics $T^2_{Obs} = U_L^2 / \sum_{i=1}^r \sigma_i^2$ using the data;
  - Compute the associated p–value $= \Pr(\chi^2_1 > T^2_{Obs})$;
  - Apply the decision rule, i.e., reject H_0 in favor of H_1 if p–value $< \alpha$.

The log–rank test

- The log–rank test set the hypothesis to

$$H_0 : t \mapsto S_A(t) - S_B(t) \equiv 0 \quad \text{vs.} \quad H_1 : t \mapsto S_A(t) - S_B(t) \not\equiv 0$$

- It consists in the following steps:
  - Fix the Type I error to some level α;
  - Compute the observed test statistics $T^2_{Obs} = U_L^2 / \sum_{i=1}^r \sigma_i^2$ using the data;
  - Compute the associated p–value $= \Pr(\chi^2_1 > T^2_{Obs})$;
  - Apply the decision rule, i.e., reject H_0 in favor of H_1 if p–value $< \alpha$.

Remark. This test is known as the log-rank test but has several names including Mantel–Haenszel, Mantel–Cox, Peto–Mantel–Haenszel.

The Wilcoxon test

- The Wilcoxon test, sometimes called Breslow test, is similar to the logrank test, i.e.,

$$H_0 : t \mapsto S_A(t) - S_B(t) \equiv 0 \quad \text{vs.} \quad H_1 : t \mapsto S_A(t) - S_B(t) \not\equiv 0,$$

  but uses a different test statistic.
- More precisely the test statistic is now $T = U_W^2 / V_W$ where

$$U_W = \sum_{i=1}^{r} n_i (d_{Ai} - \mu_{Ai}), \qquad V_W = \sum_{i=1}^{r} n_i^2 \sigma_i^2,$$

  and under the null $T \sim \chi_1^2$.
- We keep on doing the test as for that of the log–rank.

Log–rank vs. Wilcoxon

- There is a subtle difference in the test statistics of these two tests.
- To be more precise we have

$$U_L = \sum_{i=1}^{r}(d_{Ai} - \mu_{Ai}) \qquad \text{and} \qquad U_W = \sum_{i=1}^{r} n_i(d_{Ai} - \mu_{Ai})$$

- Hence the Wilcoxon approach puts less emphasis on the largest survival times (since in that case n_i are smaller)
- In practice we often use the log–rank test unless the assumption of proportional hazards (to be defined later on) is completely flawed.

Generalization of Wilcoxon

- We can embed both log–rank and Wilcoxon test statistics using weights, i.e.,

$$U_G = \sum_{i=1}^{r} w(t_{(i)})(d_{Ai} - \mu_{Ai}),$$

  where $w(t_{(i)})$ are postive weights.
- Common choices for $w(t_{(i)})$ are:
  - $w(t_{(i)}) = 1$ leading to the log–rank test
  - $w(t_{(i)}) = n_i$ leading to the Wilcoxon test
  - $w(t_{(i)}) = \hat{S}(t_{(i)})$ leading to the Peto & Peto test
  - $w(t_{(i)}) = \hat{S}(t_{(i)})^{\rho}$, $0 \leq \rho \leq 1$, leading to the Flemington–Harrington test

Extension to more than 2 groups

- In many application we may have more than 2 conditions A and B
- Fortunately both the log–rank and the Wilcoxon tests extend easily to such situations.
- Suppose that we now have $G > 2$ conditions. Similarly to what we just did, we define

$$U_{L,g} = \sum_{i=1}^{r} (d_{gi} - \mu_{gi}), \qquad U_{W,g} = \sum_{i=1}^{r} n_i (d_{g_i} - \mu_{gi}), \qquad g = 1, \ldots, G.$$

- The only difficulty is that we now have to compute covariances between $U_{L,g}$ and $U_{L,g'}$. As the formula are a bit tedious, I skip them.
- All you have to know is that the test statistic satisfies under the null $T \sim \chi^2_{G-1}$.

The `imotor` dataset

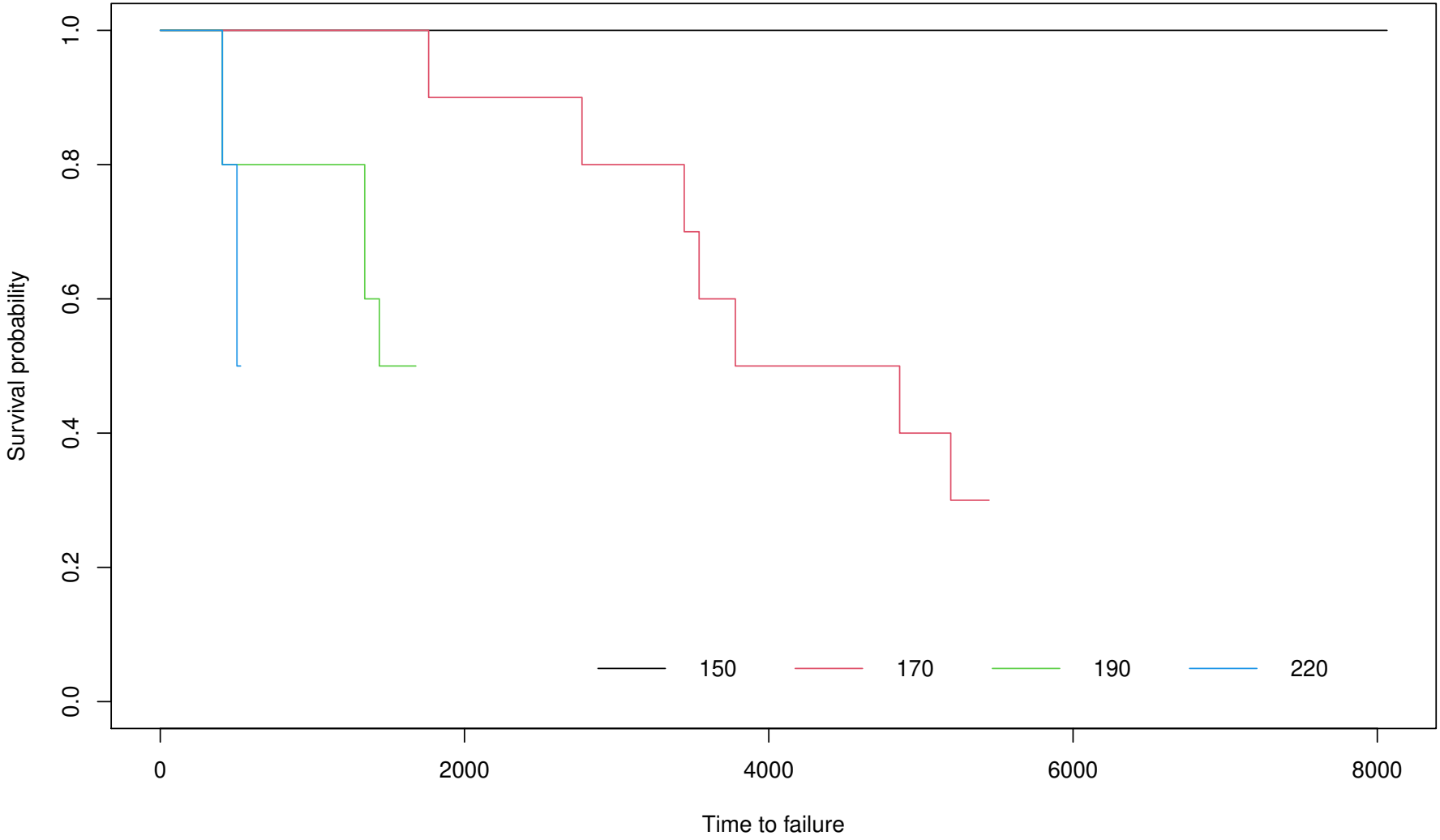


Figure 8: *Kaplan–Meier estimate of the time to failure of motor insulation at different temperatures (`imotor` dataset).*

The imotor dataset (2)

```
> survdiff(Surv(time, status) ~ temp, imotor)## <<-- log--rank

          N Observed Expected (O-E)^2/E (O-E)^2/V
temp=150 10        0     7.11     7.108    14.083
temp=170 10        7     5.76     0.269     0.431
temp=190 10        5     2.47     2.595     3.657
temp=220 10        5     1.67     6.667     9.407

 Chisq= 23  on 3 degrees of freedom, p= 4e-05

> survdiff(Surv(time, status) ~ temp, imotor, rho = 1)## <<-- Wilcoxon

          N Observed Expected (O-E)^2/E (O-E)^2/V
temp=150 10     0.00     5.21  5.212054  12.03887
temp=170 10     4.38     4.44  0.000702   0.00138
temp=190 10     4.42     2.25  2.090372   3.21890
temp=220 10     4.70     1.60  6.006250   8.83022

 Chisq= 19.9  on 3 degrees of freedom, p= 2e-04
```

Stratified test

- In some cases you may have survival times depending on some features that you assume irrelevant to the purpose of your study but may have an undesirable side effects on survival times
- To annihilate possible side effects one may use stratified tests.
- An example of stratification is for instance if in our `imotor` dataset we have a variable indicating the location of the technical centre where was conducted the experiment. Each center is thus a stratum. We may not be interested in analyzing center's behavior but rather the overall failure of motor insulation.
- Shortly stratified test just compute the same statistics (log–rank or Wilcoxon) on each stratum and add them together leading (assuming independence) to get

$$T = \frac{\sum_{s \in \text{Strata}} U_{L/W,s}^2}{\sum_{s \in \text{Strata}} V_{L/W,s}} \sim \chi_1^2$$

Lung cancer

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

```
inst:       Institution code
time:       Survival time in days
status:     censoring status 1=censored, 2=dead
age:        Age in years
sex:        Male=1 Female=2
ph.ecog:    ECOG performance score as rated by the physician.
            0=asymptomatic, 1= symptomatic but completely ambulatory, 2= in bed
            <50% of the day, 3= in bed > 50% of the day but not bedbound, 4 =
            bedbound
ph.karno:   Karnofsky performance score (bad=0-good=100) rated by physician
pat.karno:  Karnofsky performance score as rated by patient
meal.cal:   Calories consumed at meals
wt.loss:    Weight loss in last six months
```

Lung cancer (2)

```
> survdiff(Surv(time, status)~sex + strata(inst), lung)
Call:
survdiff(formula = Surv(time, status) ~ sex + strata(inst), data = lung)

n=227, 1 observation deleted due to missingness.

        N Observed Expected (O-E)^2/E (O-E)^2/V
sex=1 137      111     93.9      3.10      8.52
sex=2  90       53     70.1      4.16      8.52

 Chisq= 8.5  on 1 degrees of freedom, p= 0.004
```

- What are we testing here?
- What can you conclude?

3. Coping with covariates

- In many situations, you have access to additional information.
- Such "extra" variables are called covariates or features.[4]
- We saw some techniques on how to use such covariates, e.g., log–rank, stratification, ...
- It was a bit limited. Let's try to do better, e.g., which group of covariates impact most the survival?
- A widely used framework for this is known as the Cox proportional hazards model

[4] You can easily discriminate statisticians and non-statistician from these two words ;-)

Towards Cox proportional risk model

- Suppose we have a standard and new process for engineering something.
- Let $h_s(t)$ and $h_n(t)$ be their corresponding hazard rates.
- We assume that these hazard rates are proportional, i.e., there exists a positive constant ψ such that

$$h_n(t) = \psi h_s(t), \qquad t > 0.$$

- Two situations may arise:
  - $\psi < 1$ and the hazard rate for the new process is smaller than the standard one. improvement.
  - $\psi > 1$ and the hazard rate for the new process is larger than the standard one. standard process superior.

Cox proportional risk model

- Suppose we have n observations $T_1, \ldots, T_n$ "tied" with some additional features $\mathbf{x}_i = (x_{1,i}, \ldots, x_{p,i})^\top$, $i = 1, \ldots, n$.
- The Cox proportional risk model is defined by

$$h_i(t) = \exp(\mathbf{x}_i^\top \boldsymbol{\beta}) h_0(t),$$

where $\boldsymbol{\beta}$ is a parameter to be estimated, h_i is the hazard rate for the i–th individuals and h_0 is a baseline hazard rate.

- Note that there is no intercept since

$$\exp(\beta_0 + \mathbf{x}_i^\top \boldsymbol{\beta}) h_0(t) = \exp(\mathbf{x}_i^\top \boldsymbol{\beta}) \exp(\beta_0) h_0(t) = \exp(\mathbf{x}_i^\top \boldsymbol{\beta}) \tilde{h}_0(t),$$

i.e., intercept is included within the baseline hazard.

Different type of features

- There are two main types for the features x_j, $j = 1, \ldots, p$:
  - variates correspond to variables which take numerical values, e.g., age, number of components.
  - factors correspond to variables which have a finite number of possible outcome called levels, e.g., sex, color.

Features that are variates

- If all features are variates, then the baseline hazard h_0 corresponds to the case where the variates are all 0.
- Further, since

$$\exp(x_1\beta_1 + \cdots + (x_j + 1)\beta_j + \cdots + x_p\beta_p)h_0(t) = \exp(\beta_j)h_i(t),$$

  the $\exp(\beta_j)$ quantify the impact of one unit of increase of the j–th feature in the hazard rate when all other features being fixed to the same values.
- Consequently, three cases may arise:
  - $\exp(\beta_j) = 1$ (or $\beta_j = 0$) no effect
  - $\exp(\beta_j) > 1$ (or $\beta_j > 0$) increase in hazard
  - $\exp(\beta_j) < 1$ (or $\beta_j < 0$) decrease in hazard

Features that are factors

- Consider the case of a single feature x that is a factor with $\ell + 1$ levels $\alpha_0, \ldots, \alpha_\ell$, e.g., brown, blue, green for eyes.
- It is often most convenient to use the one hot encoding, i.e., work with the new feature

$$\tilde{x} = (1_{\{x=\alpha_0\}}, \ldots, 1_{\{x=\alpha_\ell\}}).$$

- To define appropriately the Cox model, one level has to be set as the reference level. Typically the first level is used as reference.
- Hence in this case, Cox model now writes

$$h_i(t) = \exp(\beta_1 \tilde{x}_1 + \ldots + \beta_\ell \tilde{x}_\ell) h_0(t),$$

and $\exp(\beta_j)$ quantifies how changes the hazard rate as we move from the reference level α_0 to level α_j.

Illustration on our dataset

Interaction

- Suppose now that we have two factors x_1 and x_2 (with respective levels $\alpha_j^{(1)}$ and $\alpha_k^{(2)}$)
- We may be tempted to use an interaction between these two factors, i.e., consider the impact a factor with levels $(\alpha_i^{(1)}, \alpha_j^{(2)})_{j,k}$.
- Some care is needed though, it is really an interaction effect if both factors x_1 and x_2 are used in the formulation of $h_i(t)$. We call it the main effects of x_1 and x_2.
- If only x_1 is included in the model, but not x_2, then the term $(\alpha_j^{(1)}, \alpha_k^{(2)})$ is the effect of x_2 nested within x_1
- If both x_1 and x_2 are excluded from the model, then the term $(\alpha_j^{(1)}, \alpha_k^{(2)})$ can be thought as a new factor.

Illustration on our dataset

Mixed term

- We may also be tempted to check the effect of the combination of a factor x_f and a variate x_v. Such a term is called a mixed term.
- Some care is needed in defining such mixed term, i.e., identifiability problems, but decent statistical software should manage it appropriately.

Fitting the proportional hazards model

- The model is fit by maximizing the likelihood and Sir D. Cox showed an important result, the terms $\exp(\boldsymbol{\beta}^\top \mathbf{x})$ and $h_0(t)$ can be fit separately.
- Hence if interest is only on the effect of features on the hazard rate, we need to maximize Cox's partial likelihood

$$L(\boldsymbol{\beta}) = \prod_{\substack{i:\, t_{(i)} \text{ is} \\ \text{an event}}} \frac{\exp(\boldsymbol{\beta}^\top x_i)}{\sum_{j \in R(t_{(i)})} \exp(\boldsymbol{\beta}^\top x_j)},$$

where x_i is the vector of features for the individuals who dies at time $t_{(i)}$ and $R(t_{(i)})$ is the set of indices of individuals still at risk at time $t_{(i)}$.

- There are two interesting points in this partial likelihood:
  - First individuals that are censored do not contribute to the numerator but only to the denominator;
  - Second survival times $t_{(i)}$ do not explicitely appear in $L(\boldsymbol{\beta})$ and only the ranking is needed (to compute the set $R(t_{(i)})$).

Maximization of the partial likelihood

- As usual, we actually do not maximize the partial likelihood $L(\boldsymbol{\beta})$ but minimize the negative (partial) log-likelihood

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\arg\min} \; - \log L(\boldsymbol{\beta})$$

- There is no closed form solution to this problem and numerical optimization techniques are needed.
- Typically the objective function is minimized using a Newton–Raphson algorithm which is known to converge in few iterations (for this problem).

Ties in the partial likelihood

- Cox proportional hazards model assumes that the hazard rate is continuous and in theory we cannot have ties. In practice, though, ties may occur due to rounding.
- Handling ties appropriately results in a too CPU demanding objective function, so different approximations have been suggested such as:

Breslow

$$\tilde{L}(\boldsymbol{\beta}) = \prod_{\substack{i:\, t_{(i)} \text{ is} \\ \text{an event}}} \frac{\exp(\boldsymbol{\beta}^\top s_i)}{\left\{\sum_{j \in R(t_{(i)})} \exp(\boldsymbol{\beta}^\top x_j)\right\}^{d_i}},$$

where $s_i = \sum_{j:\, t_{(j)}=t_{(i)}} x_j$ and d_i is the number of death at time $t_{(i)}$.

Ties in the partial likelihood

- Cox proportional hazards model assumes that the hazard rate is continuous and in theory we cannot have ties. In practice, though, ties may occur due to rounding.
- Handling ties appropriately results in a too CPU demanding objective function, so different approximations have been suggested such as:

  Efron

$$\tilde{L}(\boldsymbol{\beta}) = \prod_{\substack{i:\ t_{(i)} \text{ is} \\ \text{an event}}} \frac{\exp(\boldsymbol{\beta}^\top s_i)}{\prod_{k=1}^{d_i} \left\{ \sum_{j \in R(t_{(i)})} \exp(\boldsymbol{\beta}^\top x_j) - (k-1) d_i^{-1} \sum_{\ell \in D(t_{(i)})} \exp(\boldsymbol{\beta}^\top x_\ell) \right\}},$$

  where $D(t_{(i)})$ is the set of indices of individuals who died at time $t_{(i)}$.

Ties in the partial likelihood

- Cox proportional hazards model assumes that the hazard rate is continuous and in theory we cannot have ties. In practice, though, ties may occur due to rounding.
- Handling ties appropriately results in a too CPU demanding objective function, so different approximations have been suggested such as:

Cox

$$\tilde{L}(\boldsymbol{\beta}) = \prod_{\substack{i:\ t_{(i)} \text{ is} \\ \text{an event}}} \frac{\exp(\boldsymbol{\beta}^\top s_i)}{\sum_{j \in R(t_{(i)}; d_i)} \exp(\boldsymbol{\beta}^\top x_j)},$$

where $R(t_{(i)}; d_i)$ is the set of all possible draws of d_i individuals without replacement among $R(t_{(i)})$.

Model selection

- Before starting what we are about to learn is not specific to Cox's model but can be applied as long as you work with the maximum likelihood estimator.
- Model selection consists in selecting the best model among N competitive fitted models $M_1, \ldots, M_N$.
- To decide which model performs bests, we need a metric to be able to rank each of them.
- There are three main roads for this:
  - validation methods, e.g., cross-validation, data set splitting (train / validation / test)
  - Hypothesis testings, e.g., Wald test, likelihood ratio
  - Information criteria, e.g., Akaïke / Bayesian Information Criterion
- Typically we won't have huge dataset so I will focus on the last two only[5]

[5] You can find details on validation methods in other lecture I give ;-)

Asymptotic of the maximum likelihood estimator

- We start by recalling the very desirable asymptotic property[6] of the maximum likelihood estimator $\hat{\theta}$

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{\text{d.}} N\left\{0, -H(\theta_0)^{-1}\right\}, \qquad n \to \infty,$$

  where $H(\theta_0) = \mathbb{E}\{\nabla^2 \log f(X; \theta_0)\}$ where X is a single random variable from the true model $f(\cdot; \theta_0)$.
- Stating it less formally, we can argue that for n large enough

$$\hat{\theta} \dot{\sim} N\left(\theta_0, \Sigma\right),$$

  with $\Sigma = \left(-\nabla^2 \ell(\hat{\theta})\right)^{-1}$, i.e., the inverse of the Hessian matrix of the negative log-likelihood at its minimum.

[6] If regularity conditions hold...

Wald test

- The Wald test tests if a given component of the true parameter θ, say the j–th, is 0 or not, i.e.,

$$H_0 : \theta_j = 0 \qquad \text{vs.} \qquad H_1 : \theta_j \neq 0$$

- The test statistic satisfies under the null H_0

$$T_W = \frac{\hat{\theta}_j}{\text{Standard error}(\hat{\theta}_j)} \xrightarrow{\text{d.}} N(0,1), \qquad n \to \infty.$$

- We then take the decision rule as as usual.

Illustration on our data set

Likelihood ratio test

- The likelihood ratio test check wether it is preferable to opt for the nested model or the more general model.
- Recall that a model M_A is said to be nested within model M_B if M_A is a special case of M_B, e.g., the exponential distribution is nested within the Gamma family.
- The test is formalized as follows

$$H_0 : \text{nested model is correct} \qquad \text{vs.} \qquad H_1 : \text{general model is correct}$$

- The test statistic satisfies under the null H_0

$$T_{LR} = -2\left\{\ell(\hat{\theta}_{\text{gen}}) - \ell(\hat{\theta}_{\text{nest}})\right\} \xrightarrow{\text{d.}} \chi^2_p, \qquad n \to \infty,$$

where $\hat{\theta}_{\text{gen}}$ and $\hat{\theta}_{\text{nest}}$ are the maximum likelihood estimators for the general and nested models respectively and $p = \dim(\theta_{\text{gen}}) - \dim(\theta_{\text{nest}})$.

Illustration on our data set

Wald or likelihood ratio?

- The Wald and likelihood ratio tests share the same goal.
- They can be used for model selection or to test wether a feature is statistically relevant.
- They usually point to the same answer
- However for small sample size, e.g., $n \leq 30$, Wald is known to perform poorly.
- It is thus often recommended to use the likelihood ratio test.

Information criterion

- Likelihood ratio / Wald tests are useful for nested models
- If we have non nested models we can use the following information criterion:

  Akaïke $AIC(M) = -2\ell(\hat{\theta}_M) + 2\dim(\hat{\theta}_M)$
  Schwarz $BIC(M) = -2\ell(\hat{\theta}_M) + \dim(\hat{\theta}_M)\log n$,

  where $\hat{\theta}_M$ is the maximum likelihood estimator for model M.
- Both information criterion is a tradeoff between model's goodness of fit and model's complexity
- Since $\log n > 2$ when $n \geq 8$, BIC selects simpler model than AIC. Further, model selection using AIC is inconsistent.
- It is often sensible to rely on BIC

Information criterion

- Likelihood ratio / Wald tests are useful for nested models
- If we have non nested models we can use the following information criterion:

  Akaïke $AIC(M) = -2\ell(\hat{\theta}_M) + 2\dim(\hat{\theta}_M)$
  Schwarz $BIC(M) = -2\ell(\hat{\theta}_M) + \dim(\hat{\theta}_M)\log n$,

  where $\hat{\theta}_M$ is the maximum likelihood estimator for model M.
- Both information criterion is a tradeoff between model's goodness of fit and model's complexity
- Since $\log n > 2$ when $n \geq 8$, BIC selects simpler model than AIC. Further, model selection using AIC is inconsistent.
- It is often sensible to rely on BIC

☞ In practice we compute AIC and BIC for various models and select the one with the smallest information criterion value.

Model checking

- So far we learn the basic steps for the statistical modelling of survival data, i.e.,
  - a general framework, i.e., Cox's proportional hazards model
  - an estimator, i.e., the maximum (partial) likelihood estimator
  - how to get the best model from several competitive models, i.e., likelihood ratio, BIC
- However...

Model checking

- So far we learn the basic steps for the statistical modelling of survival data, i.e.,
  - a general framework, i.e., Cox's proportional hazards model
  - an estimator, i.e., the maximum (partial) likelihood estimator
  - how to get the best model from several competitive models, i.e., likelihood ratio, BIC
- However...

Model checking

- So far we learn the basic steps for the statistical modelling of survival data, i.e.,
  - a general framework, i.e., Cox's proportional hazards model
  - an estimator, i.e., the maximum (partial) likelihood estimator
  - how to get the best model from several competitive models, i.e., likelihood ratio, BIC
- However. . . you still have a poor model

Model checking

- So far we learn the basic steps for the statistical modelling of survival data, i.e.,
  - a general framework, i.e., Cox's proportional hazards model
  - an estimator, i.e., the maximum (partial) likelihood estimator
  - how to get the best model from several competitive models, i.e., likelihood ratio, BIC
- However... you still have a poor model

☞ We need to check if the best model is actually good!

Evaluating model adequacy

- Model adequacy is most often based on graphical investigation where we compare what we observed to what is expected from the fitted model: a good match indicates that the model is sensible.
- One way to do it is by analyzing residuals. Roughly speaking, a generic definition is

$$\text{residuals}_i = \text{Observation}_i - \text{Prediction}_i$$

- For instance, in the linear model $Y = X^\top \beta + \varepsilon$, we have

$$\text{residuals}_i = Y_i - X_i^\top \hat{\beta},$$

  and we usually compare those residuals to a centered Gaussian distribution.
- For the Cox's proportional model, it is a bit more tricky (in part due to censoring) and a first attempt was to analyze Cox–Snell residuals.

Cox–Snell residuals

- The motivation on the use of Cox–Snell residuals is based on the fact that

$$T_* \text{ has cumulative hazard } H \Longrightarrow H(T_*) \sim Exp(1).$$

- However since some observation may be right censored we also use the fact that

$$\mathbb{E}(H(T_*) \mid H(T_*) \geq H(C)) = H(C) + 1.$$

- Having fitted a Cox's model, it is thus sensible to define the Cox–Snell residuals as follows

$$r_{CS,i} = \begin{cases} \exp(\hat{\boldsymbol{\beta}} x_i) \hat{H}_O(T_i), & \text{if event} \\ \exp(\hat{\boldsymbol{\beta}} x_i) \hat{H}_O(T_i) + 1, & \text{if right censored,} \end{cases}$$

and to compare it to an $Exp(1)$ using a QQ–plot.

- Unfortunately, such a procedure is not very efficient in detecting departure from the fitted model.
- This is one reason they are not even implemented within the `survival` package
- To bypass this hurdle, different type of residuals have been suggested:
  - the martingale residuals
  - the deviance residuals
  - the Schoenfeld residuals
- We will cover these different types in turn.

The martingale residuals

- The martingale residuals are defined by

$$\begin{aligned} r_{M,i} &= \delta_i - r_{CS,i} \\ &= \text{observed} - \text{model prediction}, \end{aligned}$$

  where $\delta_i = 1$ if T_i is an event and 0 otherwise.
- As a consequence,
  - positive values indicate that the event actually occur sooner than what the model says
  - negative values indicate indicate that the event actually occur later than what the model says (or censored)
- In particular one can identify which individuals is poorly fit

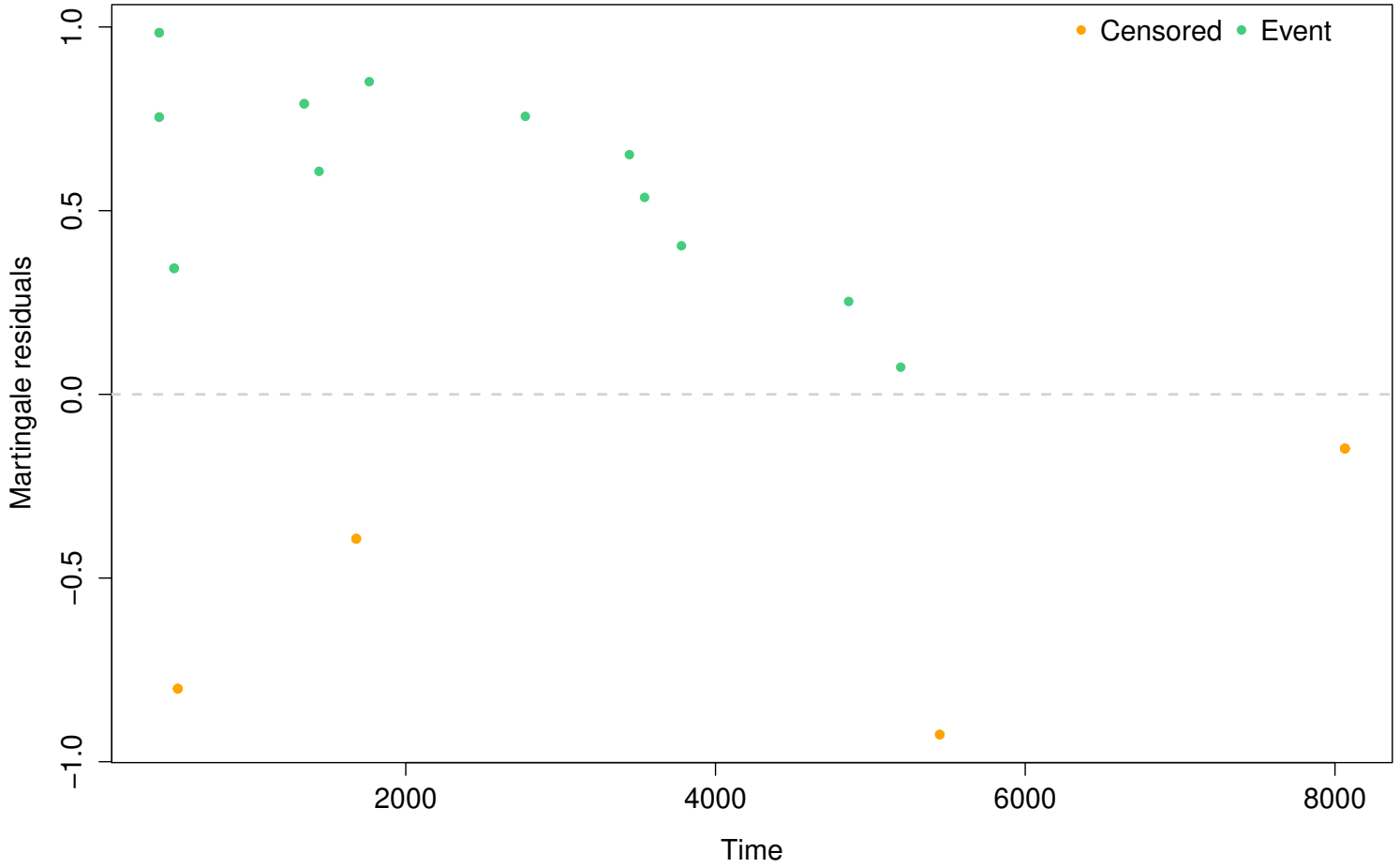


Figure 9: *Martingale residuals for the* `imotor` *data set.*

Properties of martingale residuals

- The martingale residuals have nice properties:
  - $\mathbb{E}(R_{M,i}) = 0$
  - $\text{Cov}(R_{M,i}, R_{M,j}) = 0$, for $i \neq j$
  - $\sum_{i=0}^{n} r_{M,i} = 0$
- but also undesirable ones:
  - skewed since $R_{M,i} \in (-\infty, 1]$
  - always negative for censored observation, i.e., when $\delta_i = 0$.

The deviance residuals

- The deviance residuals can be thought as a symmetrization of the martingale residuals.
- They are defined by

$$r_{D,i} = \text{sign}(r_{M,i})\sqrt{-2\left\{r_{M,i} + \delta_i \log(\delta_i - r_{M,i})\right\}},$$

and, by construction, should be approximately symmetric around 0.

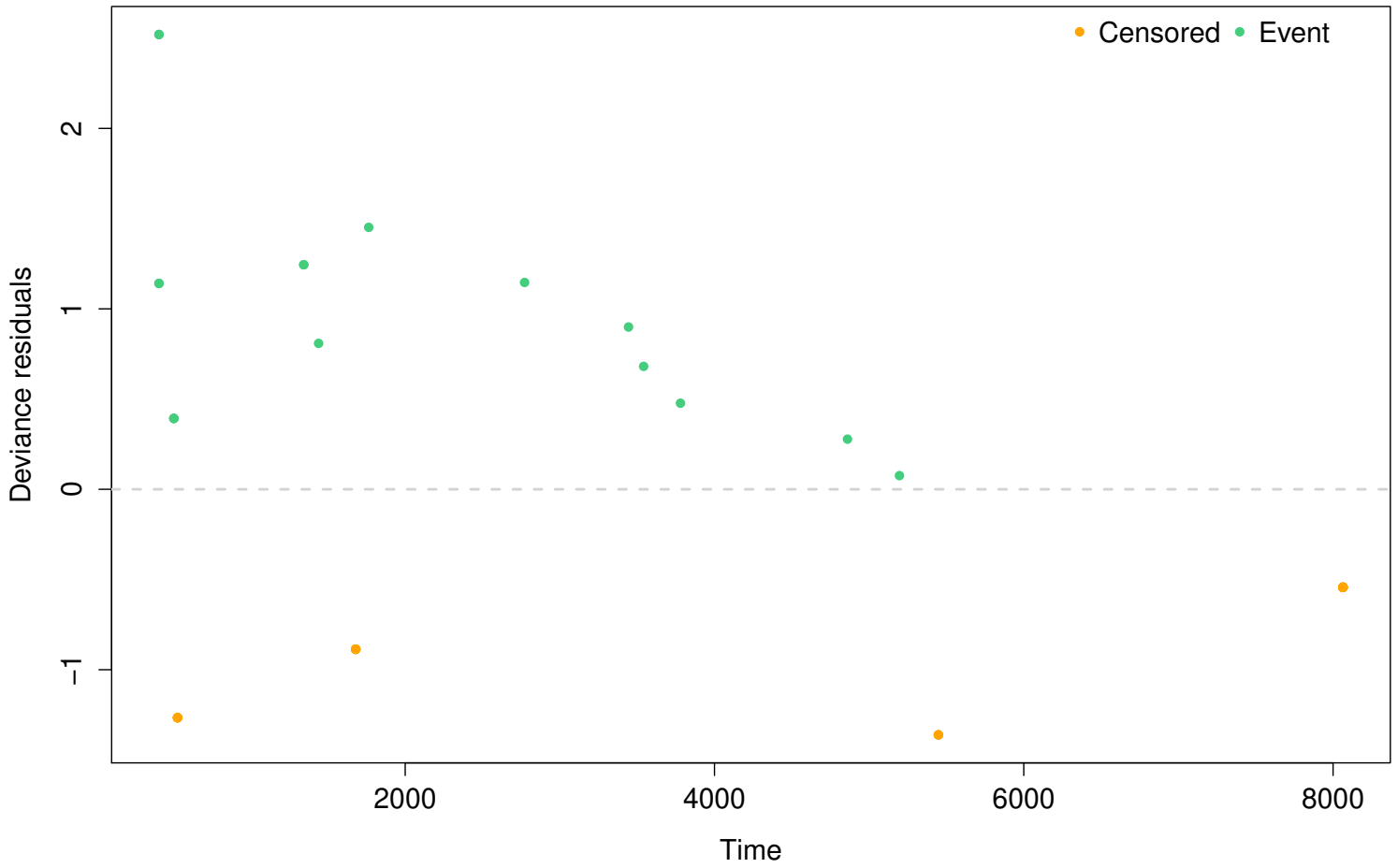


Figure 10: *Deviance residuals for the* `imotor` *data set.*

Schoenfeld residuals

- The Schoenfeld residuals differ significantly from the above residuals as, for a fixed individual, they lead to one value for each feature $x_1, \ldots, x_p$.
- They are given by

$$r_{S,ij} = \delta_i(x_{ij} - \hat{a}_{i,j}), \qquad \hat{a}_{ij} = \frac{\sum_{\ell:\, T_\ell \geq T_i} x_{\ell j} \exp(\hat{\boldsymbol{\beta}}^\top \mathbf{x}_\ell)}{\sum_{\ell:\, T_\ell \geq T_i} \exp(\hat{\boldsymbol{\beta}}^\top \mathbf{x}_\ell)}$$

- Clearly $r_{S,ij} = 0$ for censored observations (since $\delta_i = 0$) and it is common to report values only for events.
- Now if the last observation is an event then $\hat{a}_{ij} = x_{ij} \Rightarrow r_{S,ij} = 0$.
- We also have the desirable properties:
  - $\mathbb{E}(R_{S,ij}) = 0$
  - $\sum_{i=0}^{n} r_{S,ij} = 0$

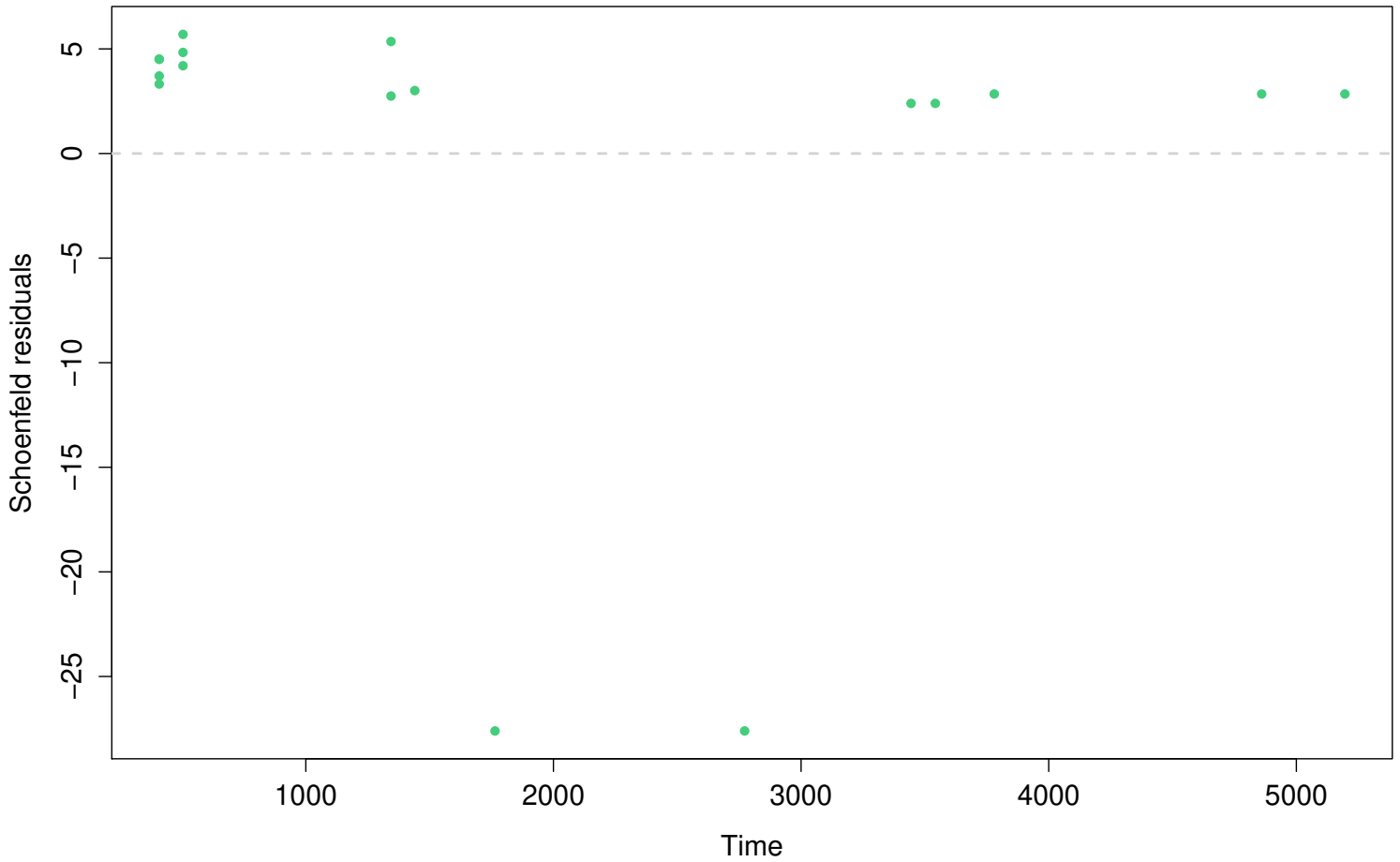


Figure 11: *Schoenfeld residuals for the* `imotor` *data set (for our single feature* `temp`*).*

Take home message about residuals

- Graphical inspection of residuals is an important stage to identify
  1. possible outliers, i.e., largest residuals in absolute values, and to understand why such a departure.
  2. possible useful transformation of the features, i.e., taking $\log temp$ rather than $temp$.
- Objective 1 is usually done by plotting residuals w.r.t. individual's index or time
- Objective 2 is usually done by plotting residuals w.r.t. to a, possibly new, covariate.

Testing proportional hazards

- In the last slides we talk about the Cox's proportional hazards model which, as its name suggests, assumes that hazards are proportional, i.e.,

$$\frac{h_i(t)}{h_j(t)} = \frac{\exp(x_i^\top \beta)}{\exp(x_j^\top \beta)}, \qquad i \neq j, \quad t > 0.$$

- What if this assumption is not supported by the data? Check!
- Since $S(t) = \exp\{-\int_0^t h(u)\mathrm{d}u\}$, we thus have

$$S_i(t) = \exp\{-\exp(x_i^\top \beta) H_0(t)\} = S_0(t)^{\exp(x_i^\top \beta)},$$

i.e., survival curves should not cross or, equivalently,

$$\log\{-\log S_i(t)\} = x_i^\top \beta + \log H_0(t),$$

i.e., the log-log survivorships should be parallel.

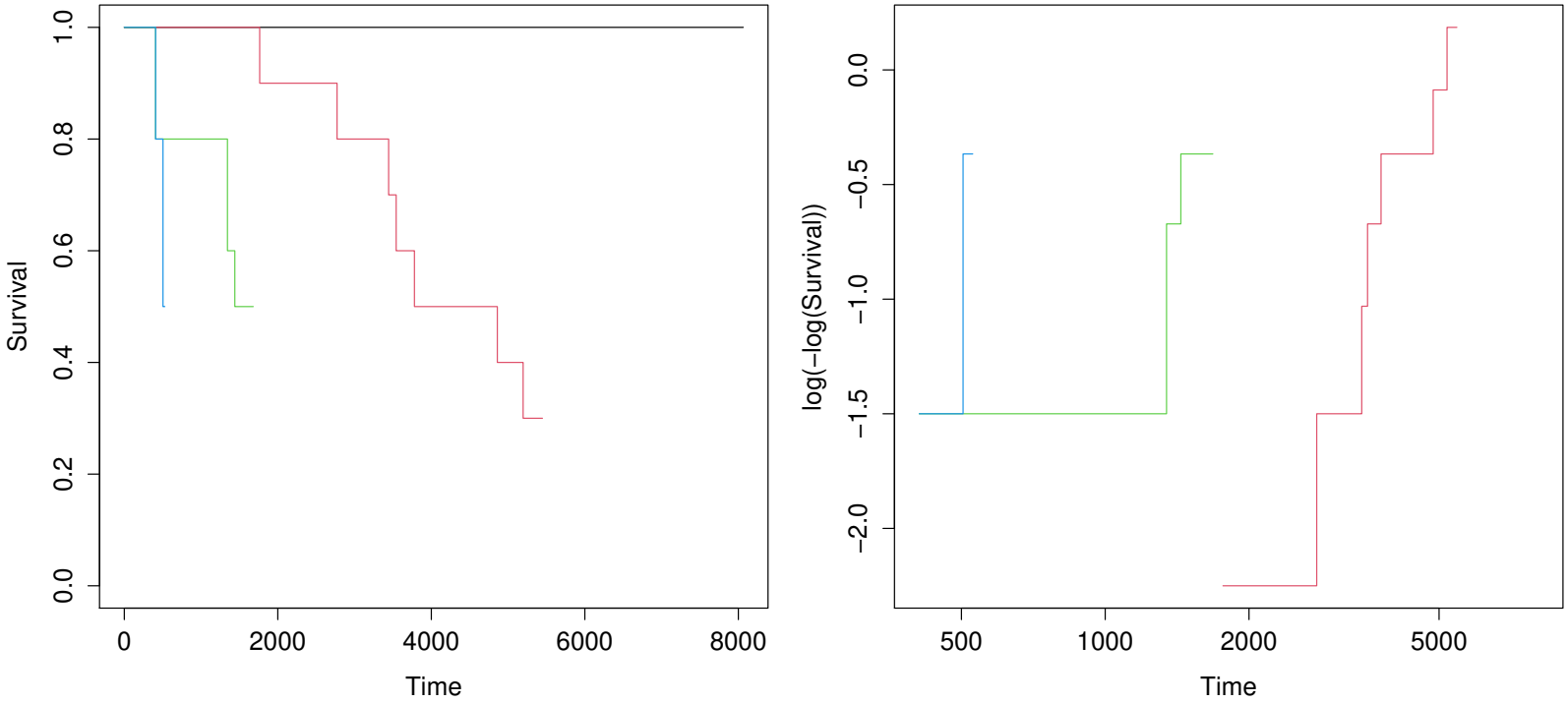


Figure 12: *Graphical assessment of the validity of the proportional hazards assumption on the* `imotor` *dataset.*

Schoenfeld residuals are back

- Consider an extended Cox's model with time–dependent coefficients, i.e.,

$$h_i(t) = h_0(t) \exp\left\{x_i^\top \boldsymbol{\beta}(t)\right\}.$$

- With such a model, the influence of a given feature, say $x_j > 0$, may vary with time: if β_j decreases with t, then x_j has less and less influence.
- Proportional hazards assumption corresponds to the case where $\boldsymbol{\beta}(t) \equiv \boldsymbol{\beta}$.
- It can be shown that

$$\beta_j(T_i) \approx \hat{\beta}_j + r_{S,ij},$$

  where $\hat{\beta}$ is the usual estimator of the Cox's model and $r_{S,ij}$ are the (scaled) Schoenfeld residuals.
- The above equation suggests to plot $\hat{\beta}_j + r_{S,ij}$ versus time (or a some function of time $g(t)$) and see whether it is constant or not

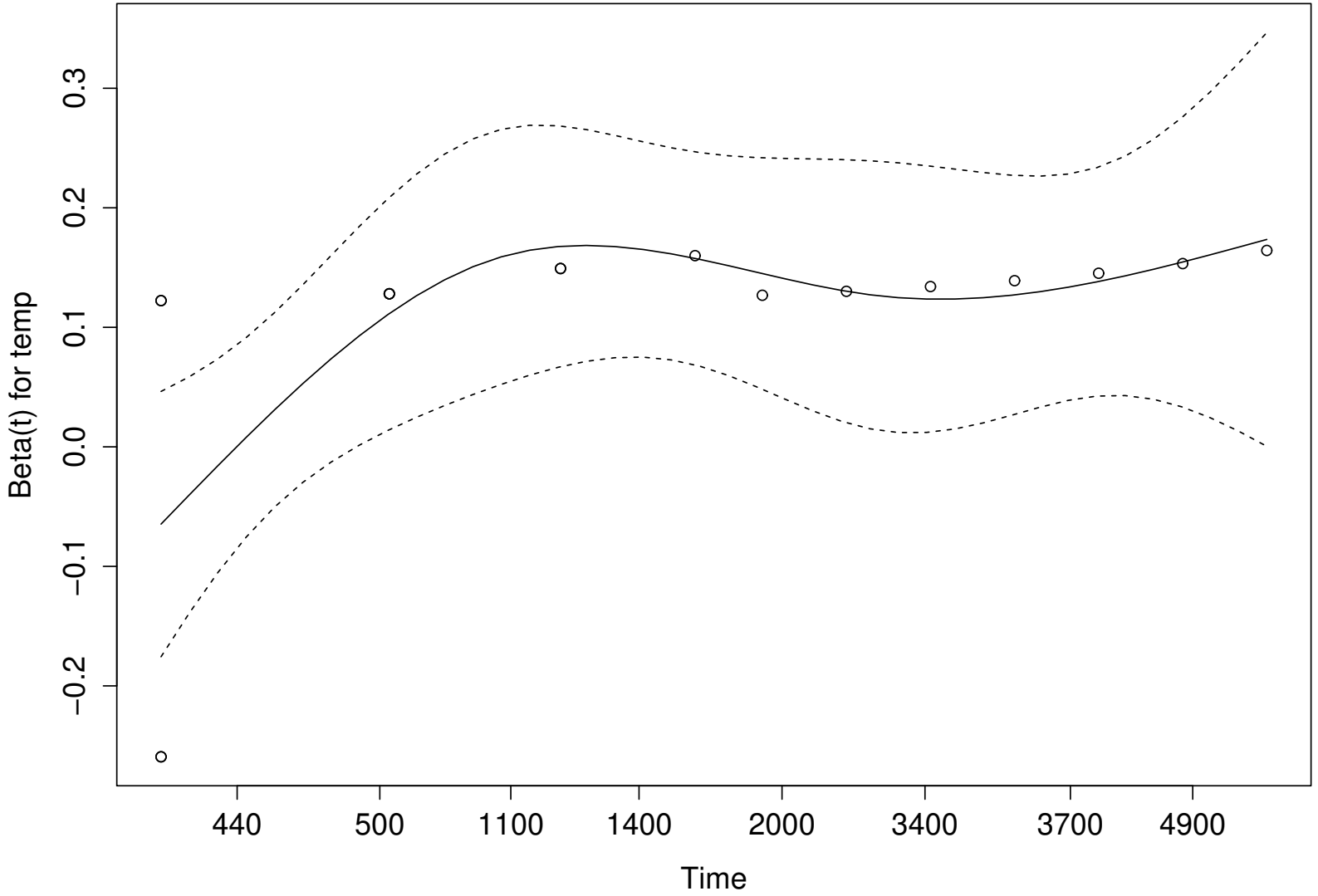


Figure 13: *Graphical assessment of the validity of the proportional hazards assumption on the* `imotor` *dataset.*

4. Time to recidivism

Time to recidivism (Rossi data set)

- Recidivism is the event re–incarceration after release from prison
- A randomized study with 52 weeks of follow–up collected information of the following variables:

`fin` financial support vs. no financial support after relase
`week` Time in week to either re–arrest or censoring
`arrest` 1 = arrest during the follow–up, 0 = no arrest
`age` Age (years) at the time of release
`race` A factor with levels `black` and `other`
`wexp` A factor with levels `yes/no` if work experience prior to incarceration
`mar` A factor with levels `(not) married` if (not) married at the time of release
`paro` A factor with levels `yes/no` if released on parole
`prio` Number of prior conviction
`educ` Categorical variable coded numerically with codes 2 (grade ≤ 6), 3 (grades 6–9), 4 (grades 10–11), 5 (grade 12) or 6 (post–secondary)

```
> head(data)
  week arrest fin age  race wexp         mar paro prio educ
1   20      1  no  27 black   no not married  yes    3    3
2   17      1  no  18 black   no not married  yes    8    4
3   25      1  no  19 other  yes not married  yes   13    3
4   52      0 yes  23 black  yes     married  yes    1    5
5   52      0  no  19 other  yes not married  yes    3    3
6   52      0  no  24 black  yes not married   no    2    4
```

```
> attach(data)
> table(fin)
 no yes
216 216
> table(race)
black other
  379    53
> table(wexp)
 no yes
185 247
> table(mar)
    married not married
         53         379
```

```
> table(paro)
 no yes
165 267
> table(educ)
  2   3   4   5   6
 24 239 119  39  11
> summary(age)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   17.0    20.0    23.0    24.6    27.0    44.0
> summary(prio)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   1.000   2.000   2.984   4.000  18.000
```

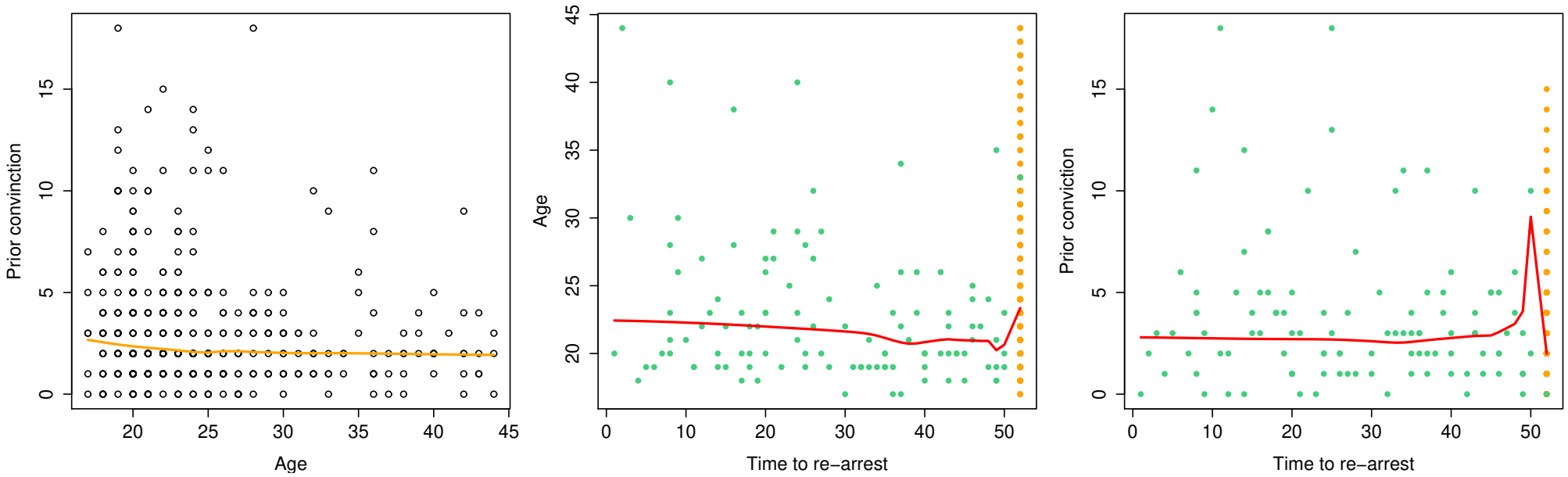


Figure 14: *Some exploratory plots for the Rossi dataset.*

- No clear trend of features w.r.t. to time to recidivism.

Probability to no recidivism

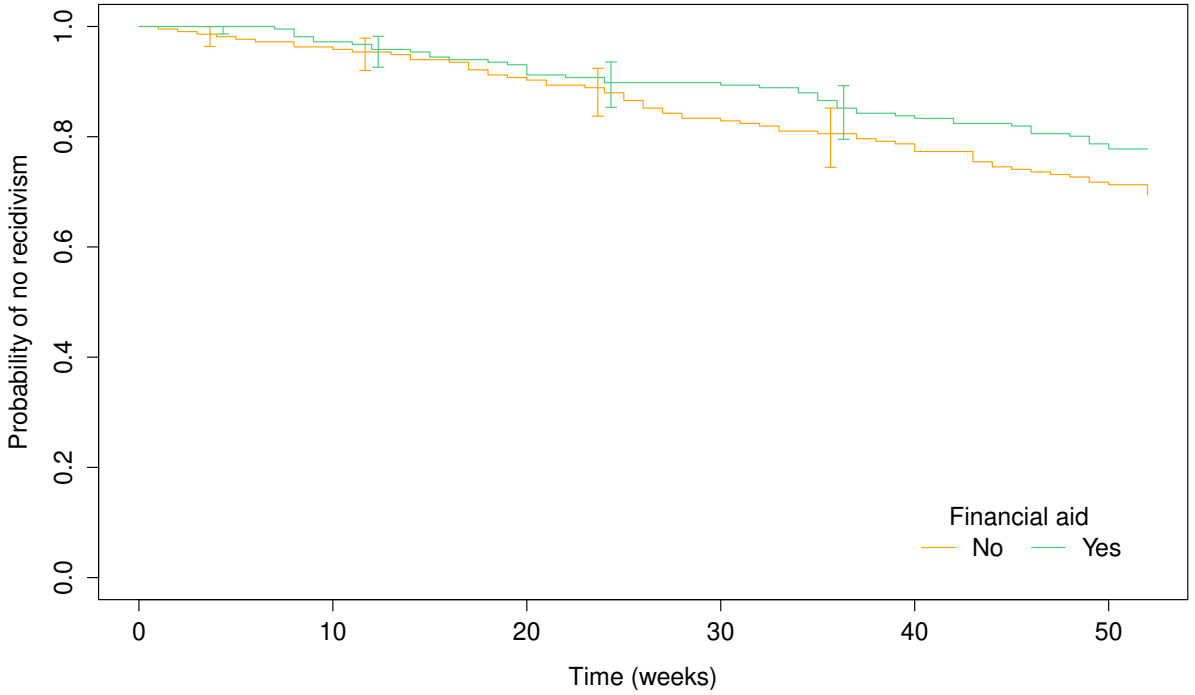


Figure 15: *Kaplan–Meier estimates according to the financial support status.*

- The curve corresponding to no financial support is always below, i.e., recidivism appears to occur sooner w/o financial aid
- This point has to be mitigated as confidence intervals overlap.
- Further other features may be related to this behavior

Is there an effect of financial support?

```
> survdiff(Surv(week,arrest)~fin, data = data)

          N Observed Expected (O-E)^2/E (O-E)^2/V
fin=no  216       66     55.6      1.96      3.84
fin=yes 216       48     58.4      1.86      3.84

 Chisq= 3.8  on 1 degrees of freedom, p= 0.05
> survdiff(Surv(week,arrest)~fin, data = data, rho = 1)

          N Observed Expected (O-E)^2/E (O-E)^2/V
fin=no  216     57.5     48.5      1.68      3.75
fin=yes 216     41.8     50.9      1.60      3.75

 Chisq= 3.7  on 1 degrees of freedom, p= 0.05
```

- The p–value is unfortunately on the boundary of the decision rule for a Type I error of $\alpha = 5\%$, i.e., no clear cut decision if an effect.
- stratified versions of the test (not shown) gives the same results (p–values ranging from 0.04 to 0.05).

Cox's regression

```
> (fit <- coxph(Surv(week,arrest) ~ fin + age + race + wexp + mar +
 paro + prio + factor(educ), data = data))

                   coef exp(coef) se(coef)      z       p
finyes          -0.40265   0.66855  0.19295 -2.087 0.03691
age             -0.05141   0.94989  0.02220 -2.316 0.02055
raceother       -0.36151   0.69663  0.31216 -1.158 0.24683
wexpyes         -0.12002   0.88690  0.21346 -0.562 0.57393
marnot married   0.42362   1.52749  0.38216  1.108 0.26765
paroyes         -0.09822   0.90645  0.19587 -0.501 0.61604
prio             0.07944   1.08268  0.02935  2.707 0.00679
factor(educ)3    0.59335   1.81004  0.51961  1.142 0.25349
factor(educ)4    0.32844   1.38880  0.54368  0.604 0.54577
factor(educ)5   -0.12098   0.88605  0.67519 -0.179 0.85780
factor(educ)6   -0.40696   0.66567  1.12333 -0.362 0.71714

Likelihood ratio test=38.68  on 11 df, p=6.013e-05
n= 432, number of events= 114
```

- Note the use of `factor` for feature educ!
- According to Wald's test, some features may be dropped
- We need to perform model selection

One example of likelihood ratio test

- According to the previous output, Wald's test suggest to drop feature `paro`
- Is it consistent with likelihood ratio test?

```
> fit2 <- update(fit, . ~ . - paro)
> anova(fit, fit2)
Analysis of Deviance Table
 Cox model: response is  Surv(week, arrest)
 Model 1: ~ fin + age + race + wexp + mar + paro + prio + factor(educ)
 Model 2: ~ fin + age + race + wexp + mar + prio + factor(educ)
   loglik  Chisq Df P(>|Chi|)
1 -656.04
2 -656.17 0.2499  1    0.6171
```

- We have a p–value very close to that of the Wald's test (as often)
- Conclusion is we can drop feature `paro`

The case of factor with more than 2 levels

- Some care is needed when deciding if one should drop `educt` or not
- Basically there are two main paths:
  - either we completely drop the feature using likelihood ratio test
  - or we try to merge some levels as shown below

```
> educ2 <- factor(educ)
> levels(educ2) <- list('2-3' = 2:3, '4-6' = 4:6)
> table(educ2)
educ2
2-3 4-6
263 169
> fit3 <- update(fit2, . ~ . - factor(educ) + factor(educ2))
> AIC(fit2, fit3)##these are *not* nested so use AIC not LRT
    df      AIC
fit2 10 1332.335
fit3  7 1329.282
```

- Here the drop in AIC is about 3 which (very!) slightly better
- Anyway, the feature educ is not significant (not shown)

- After a stage of (careful) model selection, we finally end up with

```
> summary(bestFit)
Call:
coxph(formula = Surv(week, arrest) ~ fin + age + prio, data = data)

  n= 432, number of events= 114

           coef exp(coef) se(coef)      z Pr(>|z|)
finyes -0.34695   0.70684  0.19025 -1.824 0.068197 .
age    -0.06711   0.93510  0.02085 -3.218 0.001289 **
prio    0.09689   1.10174  0.02725  3.555 0.000378 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

       exp(coef) exp(-coef) lower .95 upper .95
finyes    0.7068     1.4148    0.4868    1.0263
age       0.9351     1.0694    0.8977    0.9741
prio      1.1017     0.9077    1.0444    1.1622

Concordance= 0.63  (se = 0.027 )
Likelihood ratio test= 29.05  on 3 df,   p=2e-06
Wald test            = 27.94  on 3 df,   p=4e-06
Score (logrank) test = 29.03  on 3 df,   p=2e-06
```

Residuals analysis, i.e., 

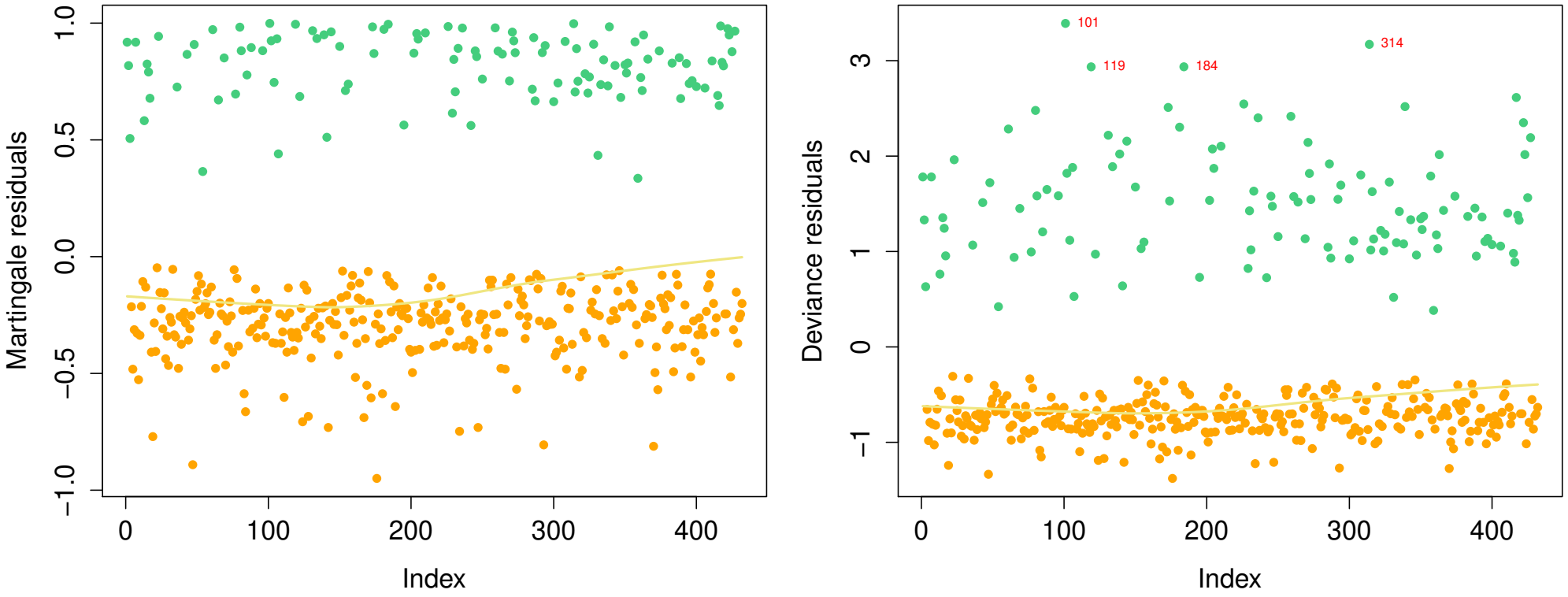


Figure 16: *Martingale and deviance residuals for our best fitted model.*

Residuals analysis, i.e., 

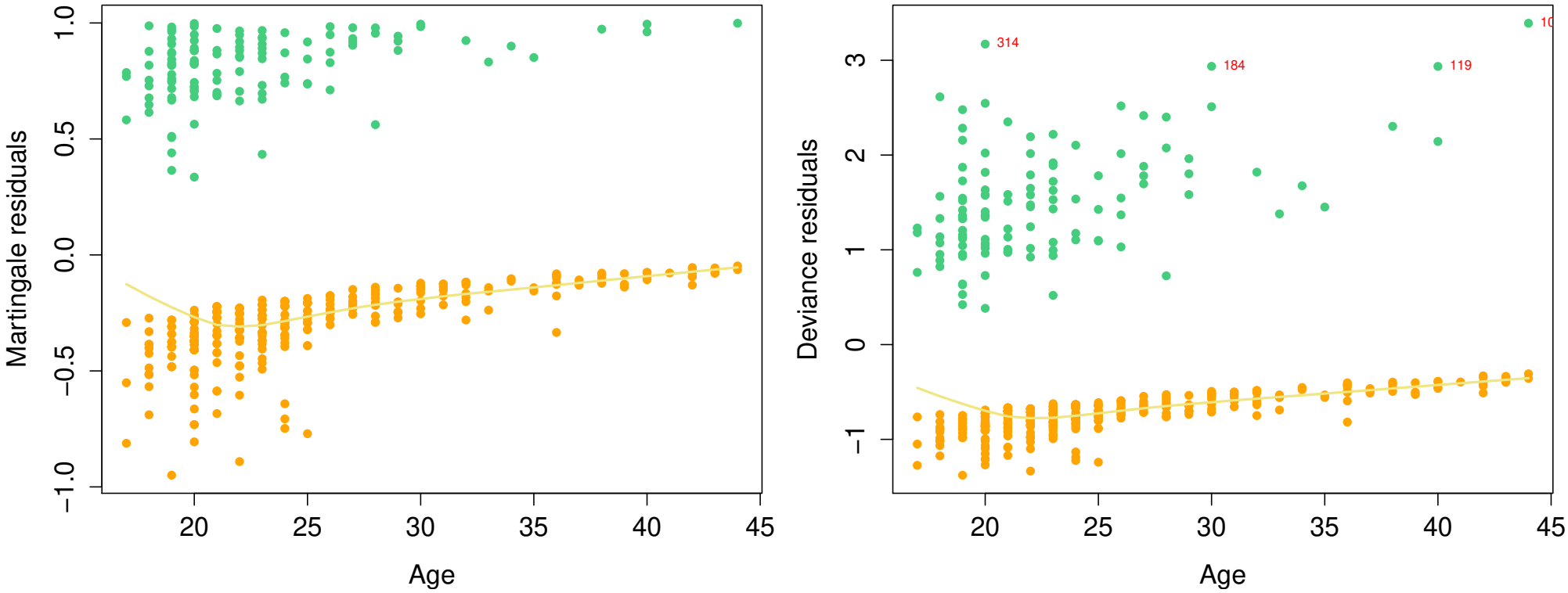
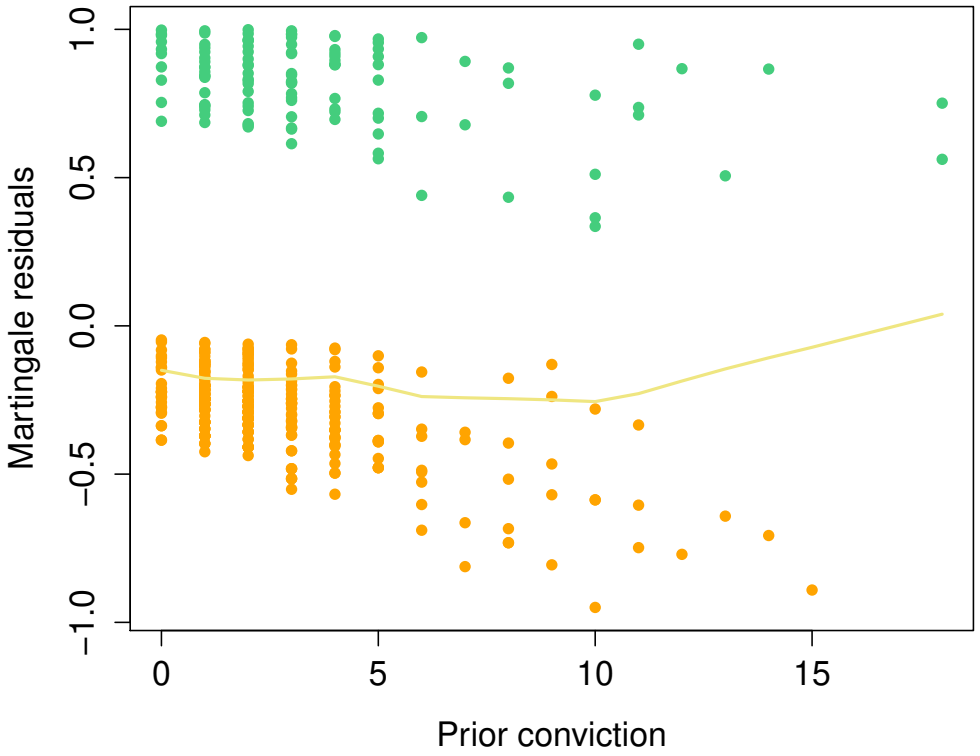


Figure 16: *Martingale and deviance residuals for our best fitted model.*

Residuals analysis, i.e., 



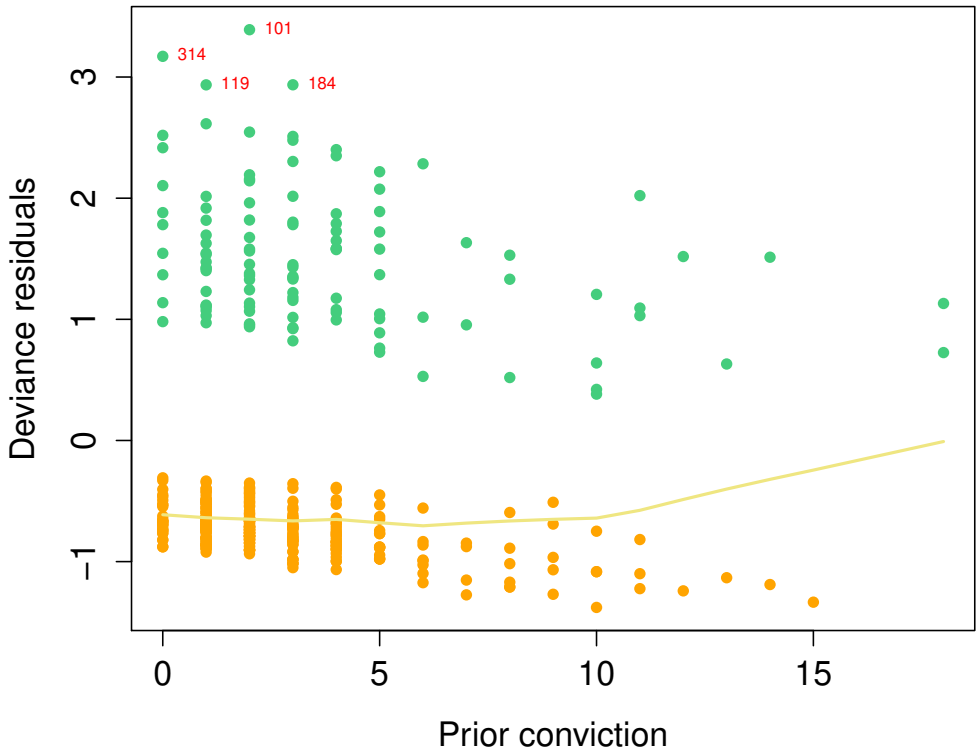


Figure 16: *Martingale and deviance residuals for our best fitted model.*

Residuals analysis, i.e., 

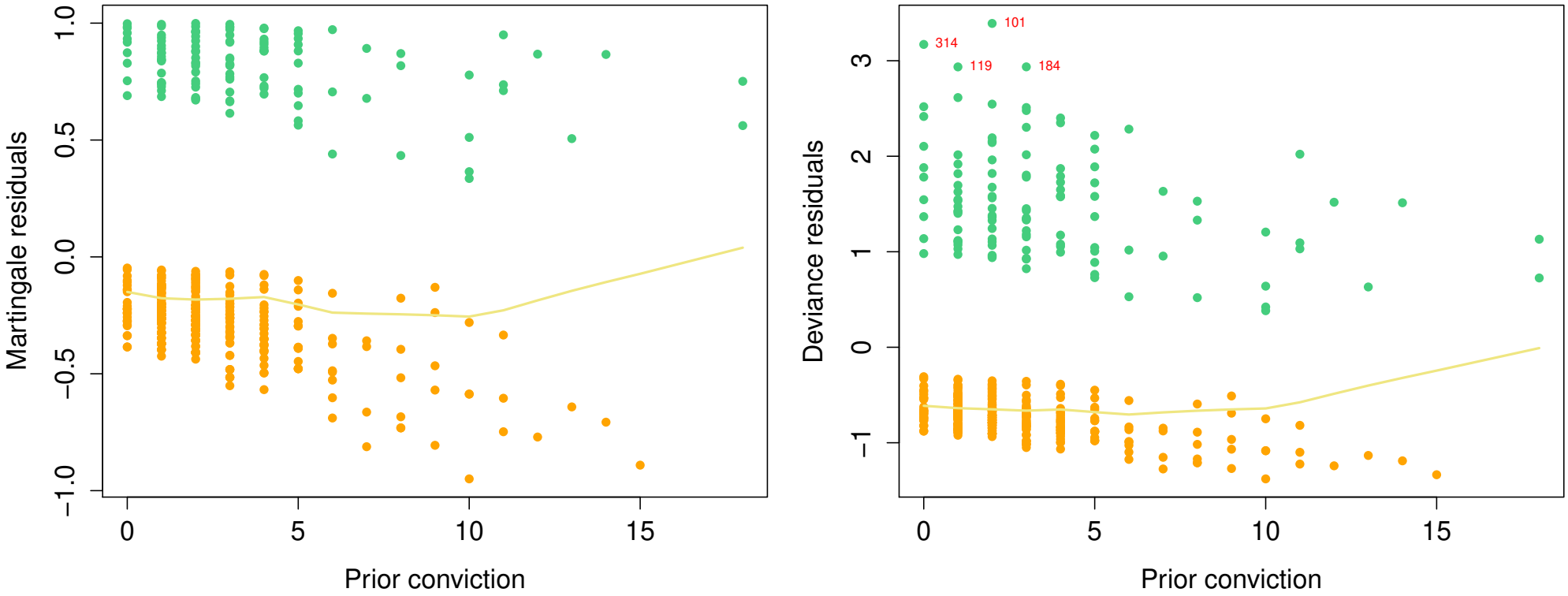


Figure 16: *Martingale and deviance residuals for our best fitted model.*

- No clear trend in residuals except for age where there is slight upward trend
- We may investigate the 4 outliers and check if influentials (not done but OK)

Outliers

```
> idx.outliers <- which(residuals(fit, "deviance") > 2.7)
> data[idx.outliers,]
    week arrest fin age  race wexp         mar paro prio educ
101    2      1  no  44 black  yes not married  yes    2    5
119    8      1 yes  40 black  yes not married  yes    1    5
184    3      1  no  30 black   no not married  yes    3    3
314    1      1  no  20 black   no not married   no    0    3
```

- They correponds to (very) early recidivism

Proportional hazards assumption checking

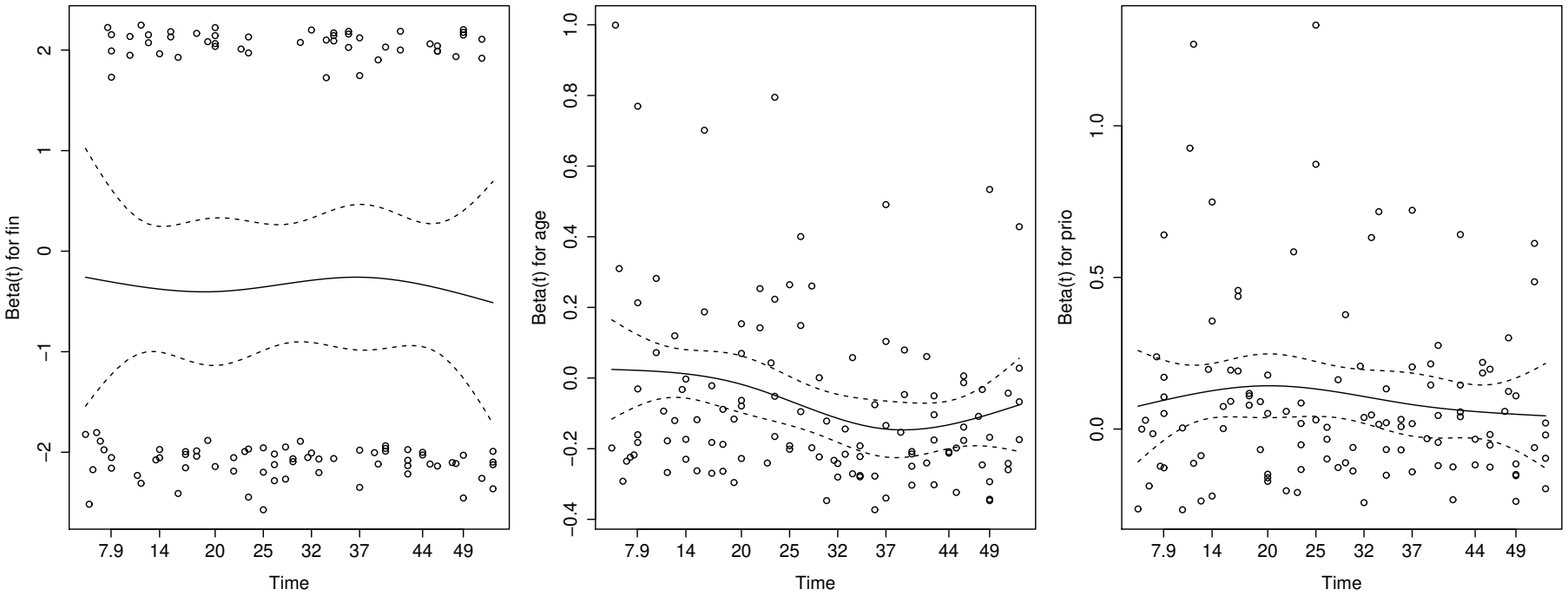


Figure 17: *Graphical assessment of the proportional hazards assumption.*

- The assumption seems OK except maybe for feature age but that's probably related to our outliers
- So we can safely state that our model is rather decent!
- We are (at last!) able to interpret our results

Interpretation (all other things being equal!)

```
> summary(bestFit)
Call:
coxph(formula = Surv(week, arrest) ~ fin + age + prio, data = data)

           coef exp(coef) se(coef)      z Pr(>|z|)
finyes -0.34695   0.70684  0.19025 -1.824 0.068197 .
age    -0.06711   0.93510  0.02085 -3.218 0.001289 **
prio    0.09689   1.10174  0.02725  3.555 0.000378 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

       exp(coef) exp(-coef) lower .95 upper .95
finyes    0.7068     1.4148    0.4868    1.0263
age       0.9351     1.0694    0.8977    0.9741
prio      1.1017     0.9077    1.0444    1.1622
```

- Financial support decreases risk by 30% (the 95% CI upper bound is slightly over 1 though)
- Being release one year older decrease the risk by 6.5% (I'm getting too old for this...)
- Having one more prior conviction increases the risk by 10%

Probability of no recidivism prediction

Figure 18: *Click on me to see the evolution of the probability of no recidivism as the number of prior conviction increases.*

Conclusion

- This lecture was pretty dense (w.r.t. the time slots allowed) but many interesting stuffs were not discussed at all.
- Of particular interest are the following:
  - time dependent features (extended Cox's model)
  - parametric survival models, e.g., Weibull distribution.
  - multiple state models, e.g., alive and tumour–free, alive and tumour present, dead.
  - frailty models, e.g., including random effects in Cox's model.

- This lecture was pretty dense (w.r.t. the time slots allowed) but many interesting stuffs were not discussed at all.
- Of particular interest are the following:
  - time dependent features (extended Cox's model)
  - parametric survival models, e.g., Weibull distribution.
  - multiple state models, e.g., alive and tumour–free, alive and tumour present, dead.
  - frailty models, e.g., including random effects in Cox's model.

THAT'S IT! I HOPE YOU ENJOYED THIS LECTURE!