Applied Biostatistics

https://moodle.epfl.ch/course/view.php?id=15590

- Review : modeling overview
- Survival data
- Censoring
- Important functions in survival analysis
- Cox regression, residuals
- Example

Modeling overview

- Want to capture important features of the *relationship between* a (set of) *variable(s)* and one or more *response(s)*
- Many models are of the form

$$g(Y) = f(\mathbf{x}) + \text{error}$$

- *Differences* in the form of g, f and distributional assumptions about the error term

Examples of models

- Linear : $Y = \beta_0 + \beta_1 x + \epsilon$
- Linear : $Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$
- (Intrinsically) Nonlinear : $Y = \alpha x_1^{\beta} x_2^{\gamma} x_3^{\delta} + \epsilon$
- Logistic regression model :

$$\log \frac{p}{1-p} = \beta_0 + \beta_1 x + \beta_2 x_2$$

- Proportional Hazards (in Survival Analysis) :

$$h(t) = h_0(t) \exp(\beta x)$$

Some background

- In logistic regression, interested in studying how risk factors are associated with presence or absence of disease (or other condition)
- Sometimes interested in how a risk factor or treatment affects *time to disease* (or some other event)
- If some study subjects *drop out*, then we may not know if they had the disease
- $\rightarrow$ *Cannot model this situation using logistic regression*

Survival data

- In many studies, an outcome of interest is the time to an event *time to an event*
- The event may be
  - *adverse* (*e.g.* death, tumor recurrence)
  - *positive* (*e.g.* leave from hospital)
  - *neutral* (*e.g.* use of birth control pills)
- Time to event data is usually referred to as *survival data* – even if the event of interest has nothing to do with 'staying alive'
- In engineering, often called *reliability data* or *failure data*

Response variable in survival analysis

- In survival analysis, the response is a *time*
- The response time $T \geq 0$
- Usually *continuous* (but measured discretely)
- One special characteristic of survival data is the presence of *censoring* – that is, *incomplete* responses
- For example, for some individuals we may know that their survival time was *at least equal* to some time t, but not know the exact time

Analysis of survival data

- *If no censoring*, could use standard regression procedures
- *However*, these may be inadequate :
  - survival time is non-negative and its distribution is generally skewed
  - *probability of surviving past a given time* typically of more interest than the *expected* time of event
  - The *hazard function*, used for regression in survival analysis, can provide more insight into the failure mechanism than linear regression
- 3 types of analyses :
  - non-parametric
  - semi-paramentric (proportional hazards)
  - (fully) parametric

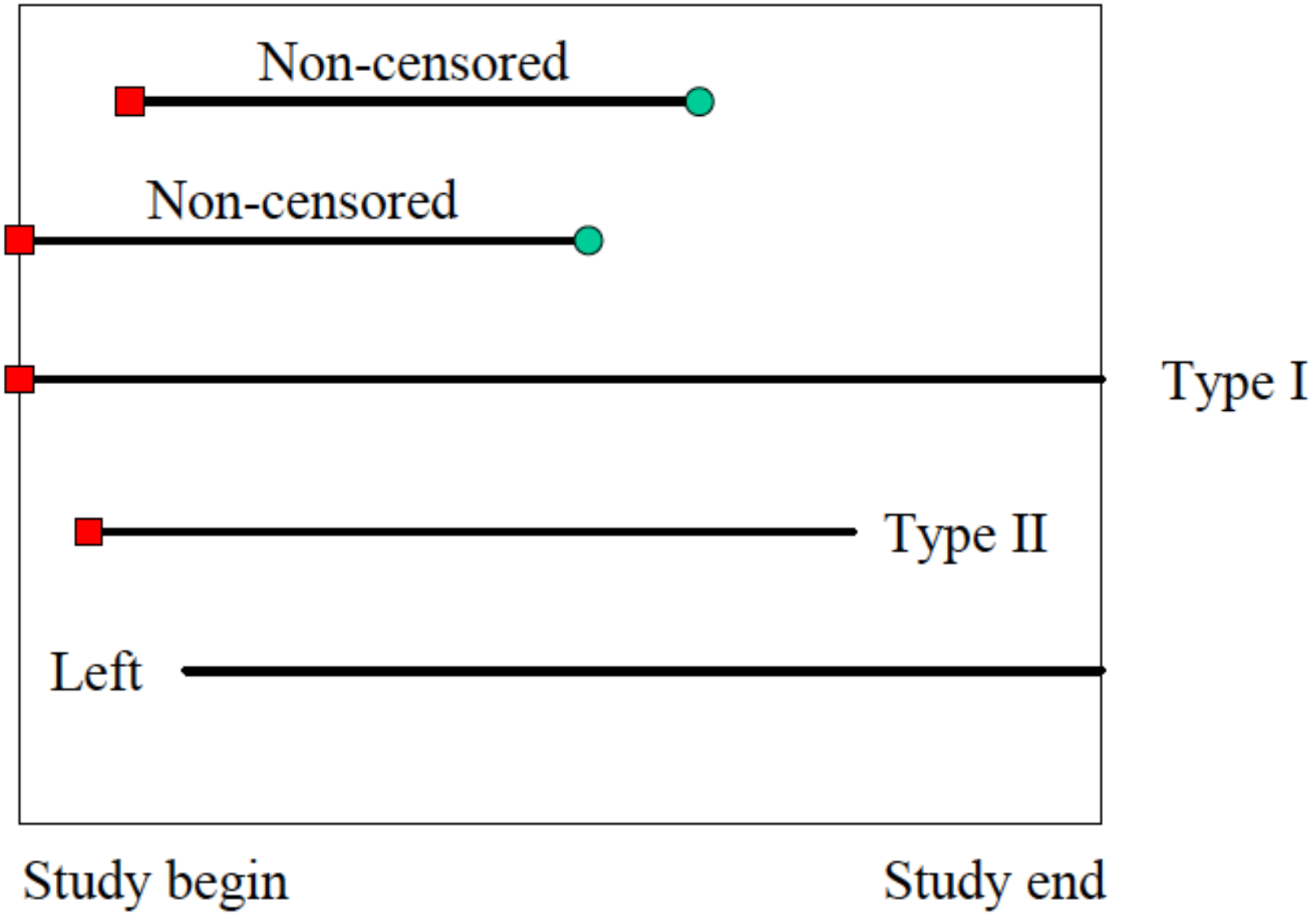
Censoring

- For some units the event of interest has occurred and therefore we know the exact waiting time, but for others it has not occurred, and all we know is that the *waiting time exceeds the observation time*
- This is called *censoring*
- For the survival analysis methods to be valid, the censoring mechanism *must be independent of the survival mechanism*
- Why censoring might occur :
  1. A subject does not experience the event before the study ends
  2. A person is lost to follow-up during the study period
  3. A person withdraws from the study
- These are all examples of *right-censoring*

Right censoring

- **Fixed type I censoring :** sample of n units is followed for a *fixed time* τ : the number of units experiencing the event (number of 'deaths') is random, but the *total duration of the study* is fixed
- **Type II censoring :**, a sample of n units is followed as long as necessary until d units (the number d is fixed in advance) have experienced the event
- More generally, for **random censoring**, each unit has associated with it a potential censoring time C_i and a potential lifetime T_i, assumed to be *independent random variables* – observe $Y_i = \min(C_i, T_i)$ and an indicator variable δ_i indicating whether observation i is terminated by death or by censoring
- For all of these schemes, the censoring mechanism is *non-informative* and they all lead to *essentially the same likelihood function*

Censoring


Non-censored
Non-censored
Type I
Type II
Left
Study begin
Study end

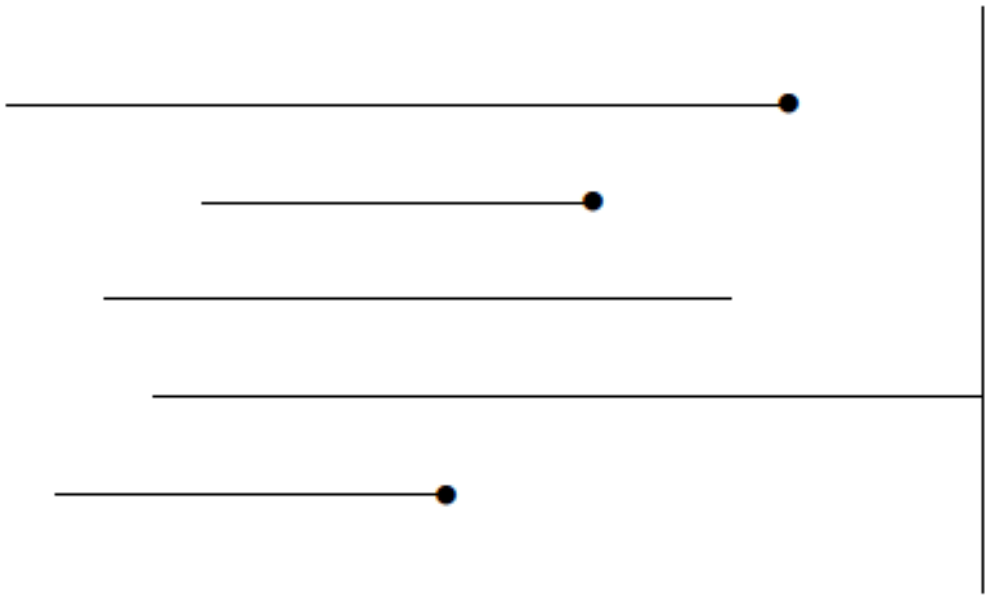
Terminology and notation

- T is the response variable, $T \geq 0$
- T_i denotes the response for the ith subject
- C_i denotes the censoring time for the ith subject
- δ_i denotes the event indicator for subject i :

$$\delta_i = \begin{cases} 1 \text{ if the event was observed } (T_i \leq C_i) \\ 0 \text{ if the event was censored } (T_i > C_i) \end{cases}$$

- $Y_i = \min(T_i, C_i)$

Example



T_i	C_i	Y_i	δ_i
80	100	80	1
40	80	40	1
74+	74	74	0
85+	85	85	0
40	95	40	1

Termination of study

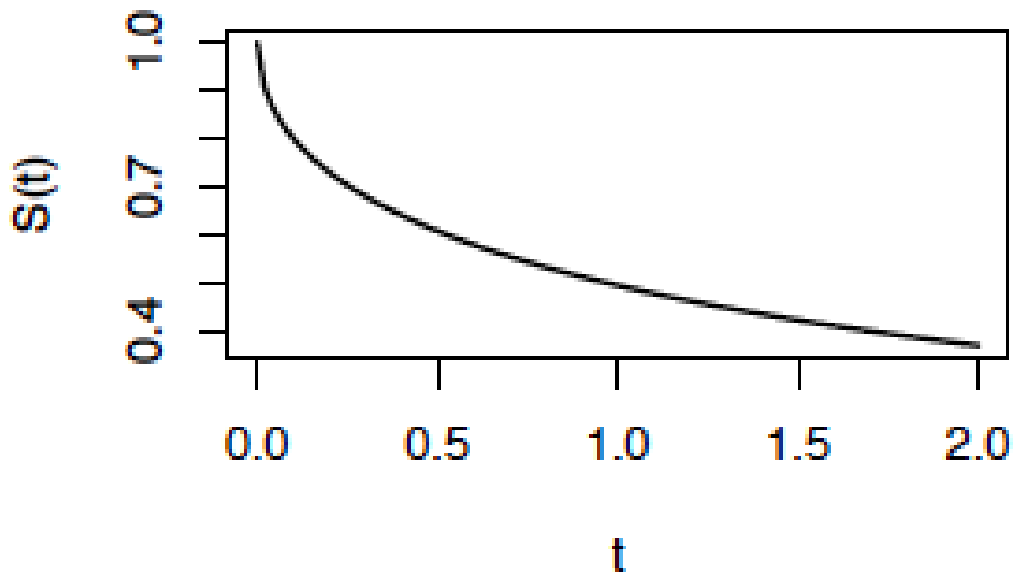
Survival function

- The *survival function* is given by :

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

where $F(t)$ is the cumulative distribution function for T

- Example :



Survival function properties

- As t ranges from 0 to ∞, the survival function has the following properties :
  - It is *non-increasing*
  - At time $t = 0$, $S(t) = 1$ (the probability of surviving past time 0 is 1)
  - At time $t \to \infty$, $S(t) = S(\infty) = 0$ (as time goes to infinity, the survival curve goes to 0 $\to$ no 'eternal life')
- In theory, the survival function is *smooth* but in practice, we observe events on a *discrete time scale* (days, weeks, *etc.*)

Estimating the survival function : no censoring (non-parametric)

- We want to estimate $S(t)$, assuming that every subject follows the same survival function (no covariates or other individual differences)
- The survival function gives the *probability that a subject will survive past time t*
- In the case of no censoring, we can use a simple non-parametric estimator of the survival function :

$$\hat{S}(t) = \frac{\text{number of individuals with survival times } \geq t}{n}$$

- This is just the (empirical) *observed proportion* of individuals surviving for at least t

Estimating the survival function : censoring

- In the presence of censoring, we cannot use this simple procedure
- In this case, typically estimate the survival function using the *Kaplan-Meier* estimator
- First, order the survival times from smallest to largest : $t_{(1)} \leq t_{(2)} \leq \cdots \leq t_{(n)}$, where $t_{(j)}$ is the jth largest unique survival time
- Kaplan-Meier estimate : $\hat{S}(t) = \prod_{j:t_{(j)} \leq t} \left(1 - \frac{d_j}{r_j}\right)$, where r_j is the number of individuals at risk just before $t_{(j)}$ (including censored individuals at $t_{(j)}$), and d_j is the number of individuals experiencing the event at time $t_{(j)}$
- For example, the survival function at the second death time $t_{(2)}$ is the estimated probability of not dying at $t_{(2)}$, conditional on the individual still being at risk at time $t_{(2)}$

More on the Kaplan-Meier estimator

- Also called the *product limit estimator*
- Typically shown graphically sometimes with confidence bands
- Has the form of a 'down staircase'
- When there is no censoring, the Kaplan-Meier curve is equivalent to the empirical distribution
- Can test for differences between curves (groups) using the *log-rank test* (below)

Variance of $\hat{S}(t)$

- In order to make *confidence intervals* for the survival function, we need to estimate the variance of $\hat{S}(t)$
- In the case of *no censoring*, $\hat{S}(t)$ is a simple proportion, so

$$Var(\hat{S}(t)) = \frac{\hat{S}(t)(1 - \hat{S}(t))}{n}$$

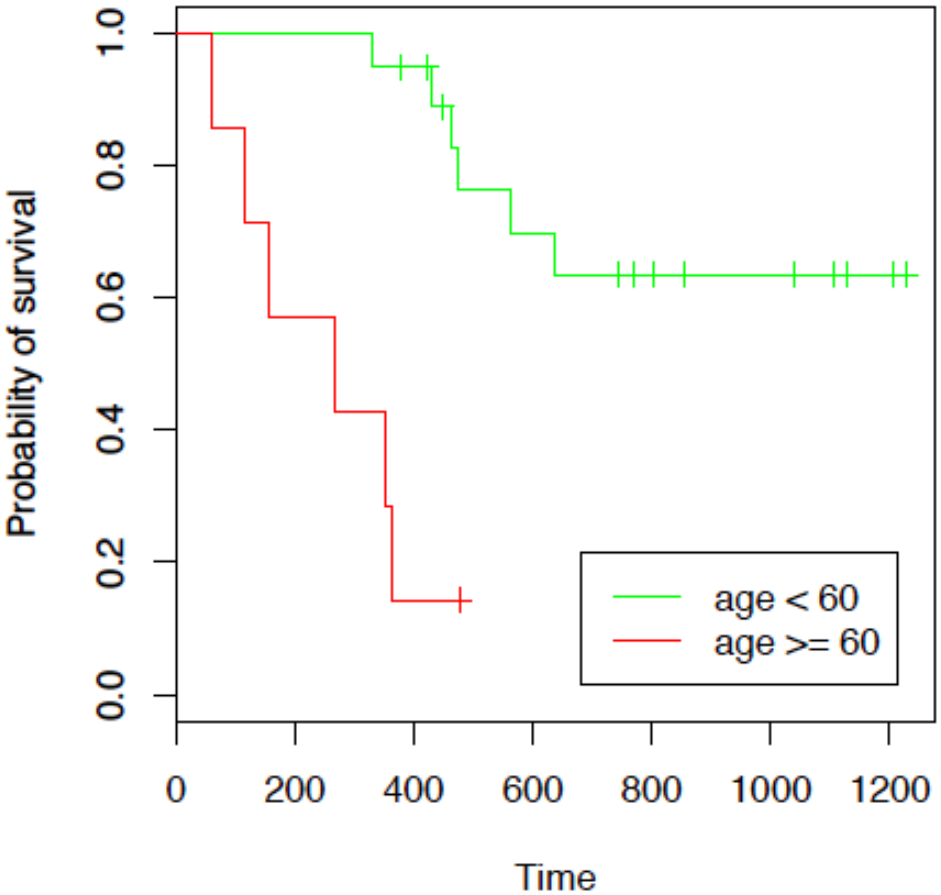
- For the *Kaplan-Meier estimate* of $\hat{S}(t)$, the variance is more difficult to derive, but it turns out to be

$$Var(\hat{S}(t)) = \left(\hat{S}(t)\right)^2 \sum_{j:t_{(j)} \leq t} \frac{d_j}{r_j \left(r_j - d_j\right)}$$

log-rank test

- We can carry out a formal hypothesis test of equality of survival curves for two groups using the *log-rank test* :
  - Compute *expected number of deaths* for each unique death time in the data, assuming that the chance of dying for subjects at risk is the same for each group
  - Total number of expected deaths for each group is the sum of the expected numbers for each time
  - The test compares the *observed* number of deaths in each group to the *expected* number using a χ^2 test
- (more on χ^2 tests after the break)

Example



- Are these two curves statistically different ?
- use of the logrank test : p-value = 3.56e-05

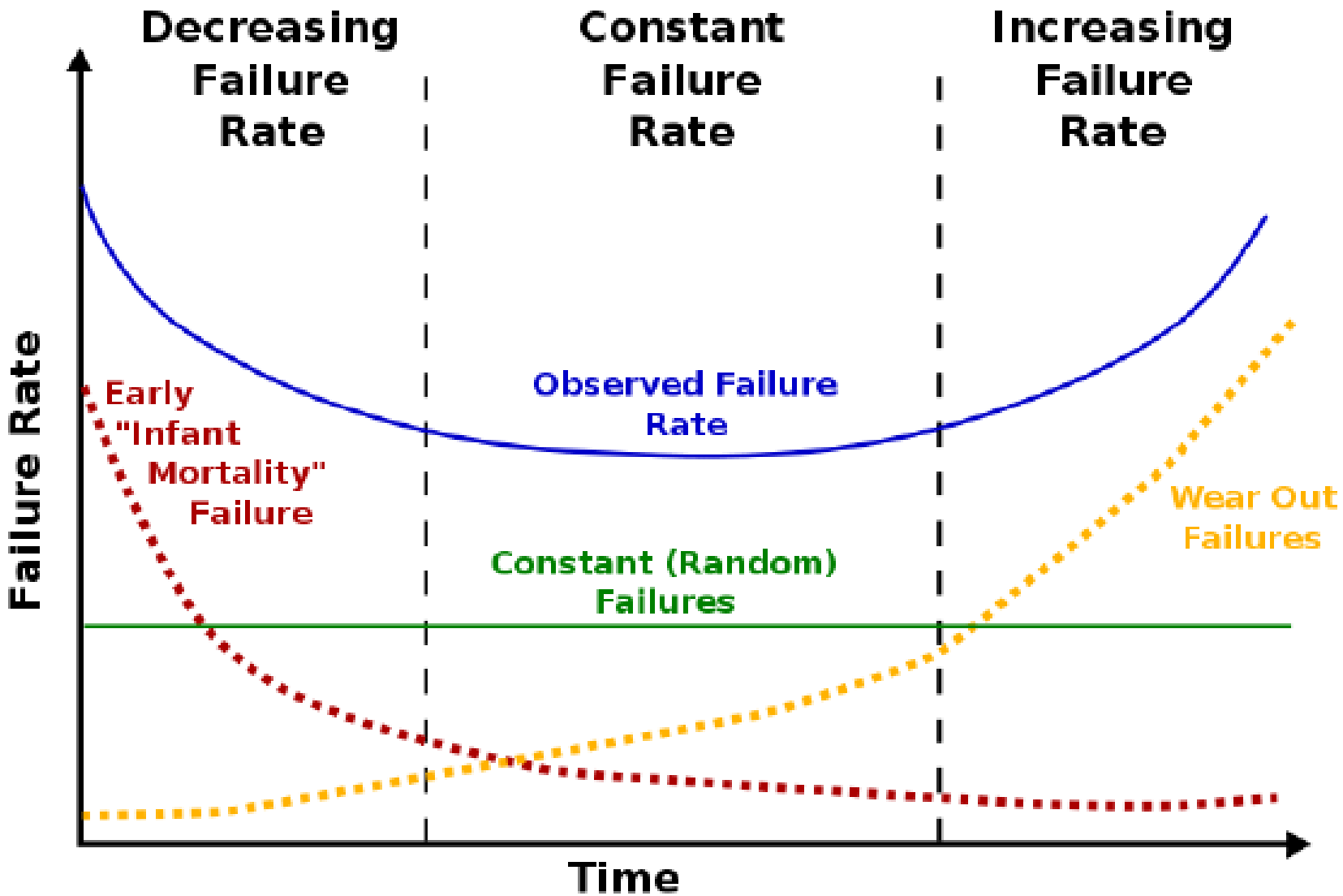
Hazard function

- Often want to assess which time periods have high or low chances of failure, among those still at risk at the time
- We can characterize these risks using the instantaneous failure rate, or *hazard function* $h(t)$
- $h(t)$ is the probability that an individual experiences the event in a small time interval s, *given that the individual has survived up to the beginning of the interval*, when the size of the time interval approaches 0 :

$$h(t) = \lim_{s \to 0} \frac{P(t \leq T \leq t + s \mid T \geq t)}{s},$$

where T is the individual's survival time

Hazard function examples



Relations between functions

- The hazard function and survivor function are related as follows :

$$S(t) = \exp\left(-H(t)\right),$$

where $H(t)$ is the *integrated hazard* or *cumulative hazard* function, defined as :

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

- Therefore, $H(t) = -\log S(t)$
- $h(t) = f(t)/S(t)$, where $f(t)$ is the *density function* for T

Estimating h, H

- We can estimate the hazard function as the proportin of individuals experiencing the event in an interval per unit time, given that they have survived to the beginning of the interval :

$$\hat{h}(t) = \frac{d_j}{n_j(t_{(j+1)} - t_{(j)})}$$

- To estimate the integrated hazard function :

$$\widehat{H}(t) = \sum_j \frac{d_j}{n_j}$$

Parametric estimation of S, H

- The estimators above for $S(t)$ and $H(t)$ are *non-parametric* – that is, we do not make any assumptions about the distributional form for the survival time T
- We could instead make *parametric assumptions* regarding the form for the distribution of T
- Some common parametric models for survival data :
  - exponential
  - Weibull
  - gamma
  - log-normal
- Assuming that the parametric model is correct, we can estimate $S(t)$ more precisely
- Estimation of parameters is by maximum likelihood

Cox regression (semi-parametric)

- Need special regression techniques to deal with (possibly) censored survival data
- Most commonly used procedure is the *Cox proportional hazards model* or *Cox regression*
- In this case, *model the hazard function*
- Cox modeling is carried out in a similar manner to regression modeling, but with linearity assumed on the *log hazard scale*
- The Cox proportional hazards model can be written as :

$$h(t) = h_0(t) \exp\left(\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k\right),$$

  where h_0 is the *baseline hazard function*, *i.e.*, the hazard function for individuals with all explanatory variables $= 0$
- This model forces the *hazard ratio* between two individuals to be constant over time

Model components and interpretation

- In the Cox model, $h_0(t)$ describes the *common shape of the survival time distribution* for all individuals
- The *relative risk function* $\exp(\beta'\mathbf{x})$ gives the level of each individual's hazard
- Model parameters typically estimated using *partial likelihood*, which takes care of the issue that the nuisance parameter h_0 is left *unspecified*
- Interpretation of parameter β_j : $\exp(\beta_j)$ gives the **relative risk change** (hazard ratio) associated with an increase of one unit in covariate x_j, all other explantory variables remaining constant
  - HR $= 1$: No effect
  - HR > 1 : Increase in hazard
  - HR < 1 : Reduction in hazard (protective)

Survival analysis in R

- `R` package `survival`
- A *survival object* is made with the function `Surv()`
- What you have to tell `Surv` :
  - *time* : observed survival time
  - *event* : indicator saying whether the event occurred (`event=TRUE`) or is censored (`event=FALSE`)
- Analyze with Kaplan-Meier curve : `survfit`
- log-rank test : `survdiff`
- Cox proportional hazards model : `coxph`

- *Annotated example :* `http://www.sthda.com/english/wiki/cox-proportional-hazards-model`

Cox PH model (again)

- The Cox (PH) model : $\lambda(t \mid \mathbf{X(t)}) = \lambda_{\mathbf{0}}(\mathbf{t})\mathbf{exp}\{\beta' \mathbf{X(t)}\}$
- Assumptions of this model :
  1. the regression effect β is constant over time (PH assumption)
  2. linear combination of the covariates (including possibly higher order terms, interactions)
  3. the link function is exponential
- The PH assumption (1) has received most attention in both research and application

Cox PH model assessment

- Typically examine different kinds of *residuals*, but not straightforward to define residuals for binary outcomes (death or not)
- Possibilities include
  1. Generalized (Cox-Snell)
  2. Schoenfeld (or weighted Schoenfeld)
  3. Martingale
  4. Deviance
- Generalized residual procedure not very sensitive for checking the Cox model

Schoenfeld residuals

- Instead of a single residual for each individual, there is a *separate residual for each individual for each covariate*
- These represent the difference between the *observed covariate* and the *expected*, *given* the risk set at that time
- Calculated for each covariate
- Not defined for *censored* failure times
- Sum of the Schoenfeld residuals $= 0$
- asymptotically uncorrelated with expectation zero under the Cox model
- $\Rightarrow$ plot of r_{ij} versus X_i should be centered around 0
- non-PH could be revealed in such a plot
- Weighted Schoenfeld residuals : weighted by the variance-covariance matrix
- Might be more Normally distributed for binary variables

Schoenfeld residuals in R

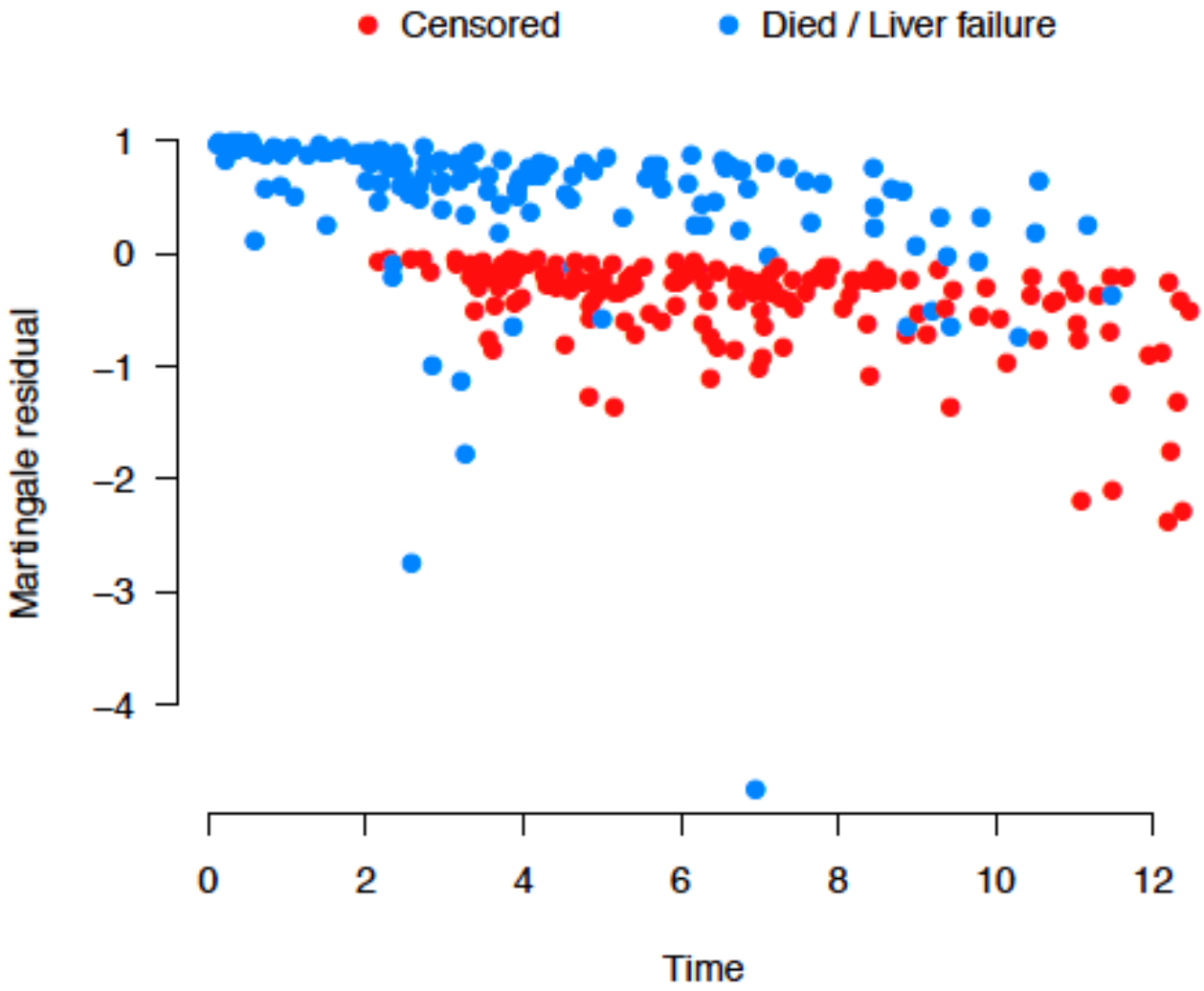
- Say you have a dataset **foo**
- R commands :

```
cox ← coxph(Surv(days,1-censor) ~ trmt,foo)
residuals(cox)
residuals(cox,type="scaledsch")
print(cox.zph(cox))
```

Martingale residuals

- The martingale residual for an observation is defined as : $\hat{m}_i = d_i - \hat{e}_i$
- That is, the discrepancy between the observed value of a subject's failure indicator and its expected value, integrated over the time for which that patient was at risk
- Positive values mean that the patient died *sooner* than expected (according to the model) ; negative values mean that the patient lived *longer than expected (or were censored)*
- In R, you can get the martingale residuals from the `survival` package by calling `residuals(fit)`, where `fit` is a fitted coxph model (`resid(fit)` also works as a shortcut)
- Several residual options, martingale residuals are returned by default

Example



The large outlier is a patient with stage 4 cirrhosis and a bilirubin concentration of 14.4 (96th percentile), yet survived 7 years

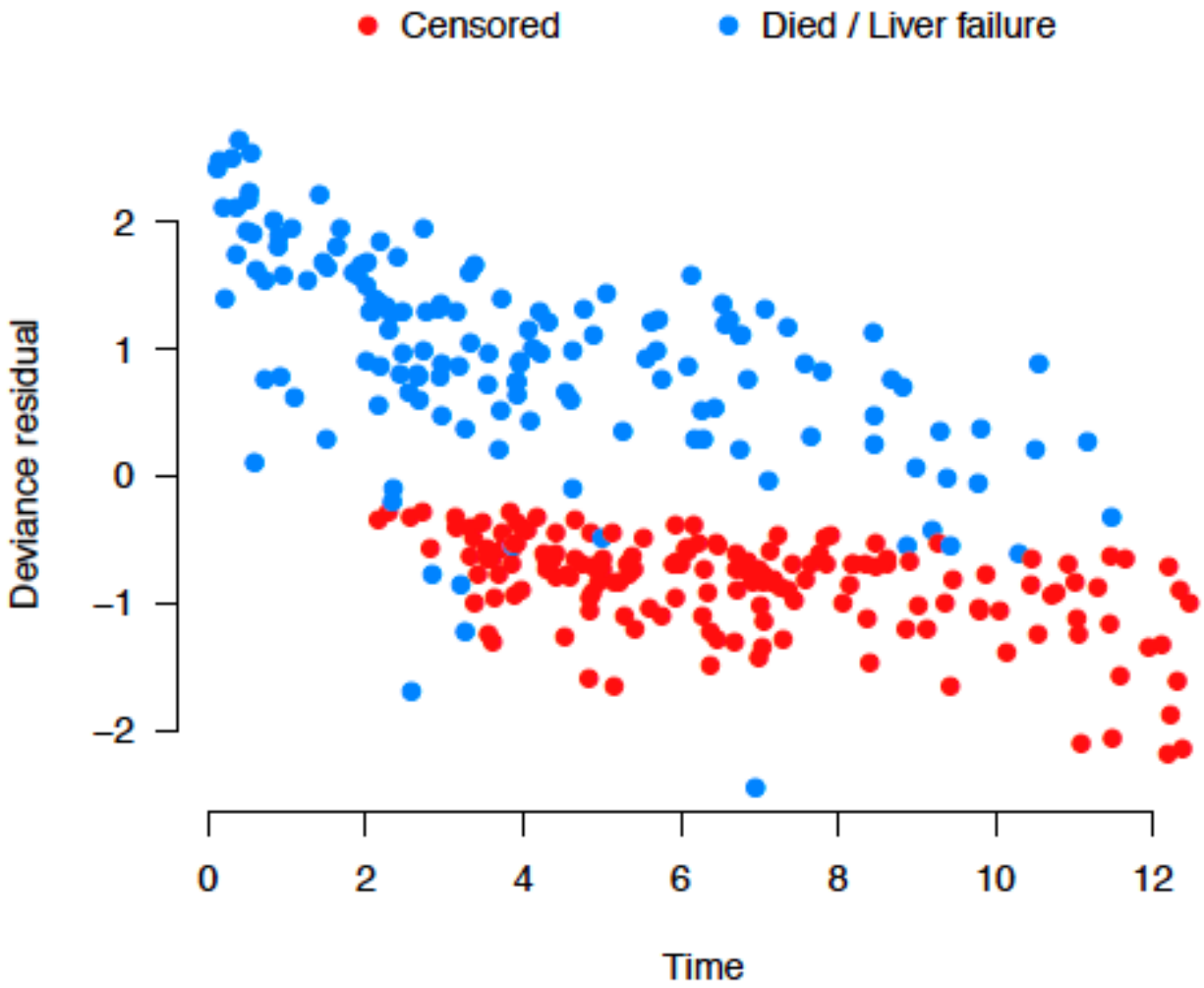
Comments on martingale residuals

- Martingale residuals are very useful and can be used for many of the usual purposes that we use residuals for in other models (identifying outliers, choosing a functional form for the covariate, etc.)
- However, the primary drawback to the martingale residual is its clear asymmetry (its upper bound is 1, but it has no lower bound)

Deviance residuals

- A technique for creating symmetric, normalized residuals that is widely used in generalized linear modeling is to construct a *deviance residual*
- The idea behind the deviance residual is to examine the difference between the log-likelihood (ℓ_i) for subject i under a given model and the maximum possible log-likelihood for that subject ($\tilde{\ell}_i$)
- As it is essentially a likelihood ratio test, the quantity $2(\tilde{\ell}_i - \ell_i)$ should approximately follow a χ^2_1 distribution
- The deviance is then defined as : $\hat{d}_i = \text{sign}(\hat{m}_i)\sqrt{2(\tilde{\ell}_i - \ell_i)}$
- Can be used to look for outliers, or plot them against covariates to assess the relationship between a covariate and unexplained variation
- Can also assess whether the relationship between the predictor and the (log) hazard is linear
- In R : `residuals(fit, type="deviance")`

Example



The deviance residuals are much more symmetric

Assessing the PH assumption

- **Graphical :**
  - Plots of survival estimates for two subgroups ; (Indications of non-PH : estimated survival curves are fairly separated, then converge or cross)
  - Plots of $\log[-\log(\hat{S}(t))]$ for two subgroups ; (if unparallel, non-PH
  - Plots of (weighted, cumulative) Schoenfeld residuals vs time ; (without cumulating : increase or decrease over time, may fit a OLS regression line to see the trend)
  - Could plot observed survival probabilities (estimated using KM) versus expected under PH model, but survival curves tend not to be sensitive

- **Formal Goodness-of-fit tests**

What can we do if PH fails ?

- *Transformations* of the covariates
- Add in *higher order terms, interactions* between covariates
- Carry out a *stratified* analysis
- Fit a *time-varying coefficients* model
- Try other models

R examples

- https://www.r-bloggers.com/cox-model-assumptions/
- http://www.sthda.com/english/wiki/cox-model-assumptions
- https://stat.ethz.ch/R-manual/R-devel/library/survival/html/residuals.coxph.html
- https://stat.ethz.ch/R-manual/R-devel/library/survival/html/plot.cox.zph.html