# BIOS 6312: Modern Biostatistics Methodology II

**Andrew J. Spieker, Ph.D.**

Associate Professor of Biostatistics
Vanderbilt University

Set 9: Analysis of time-to-event outcomes

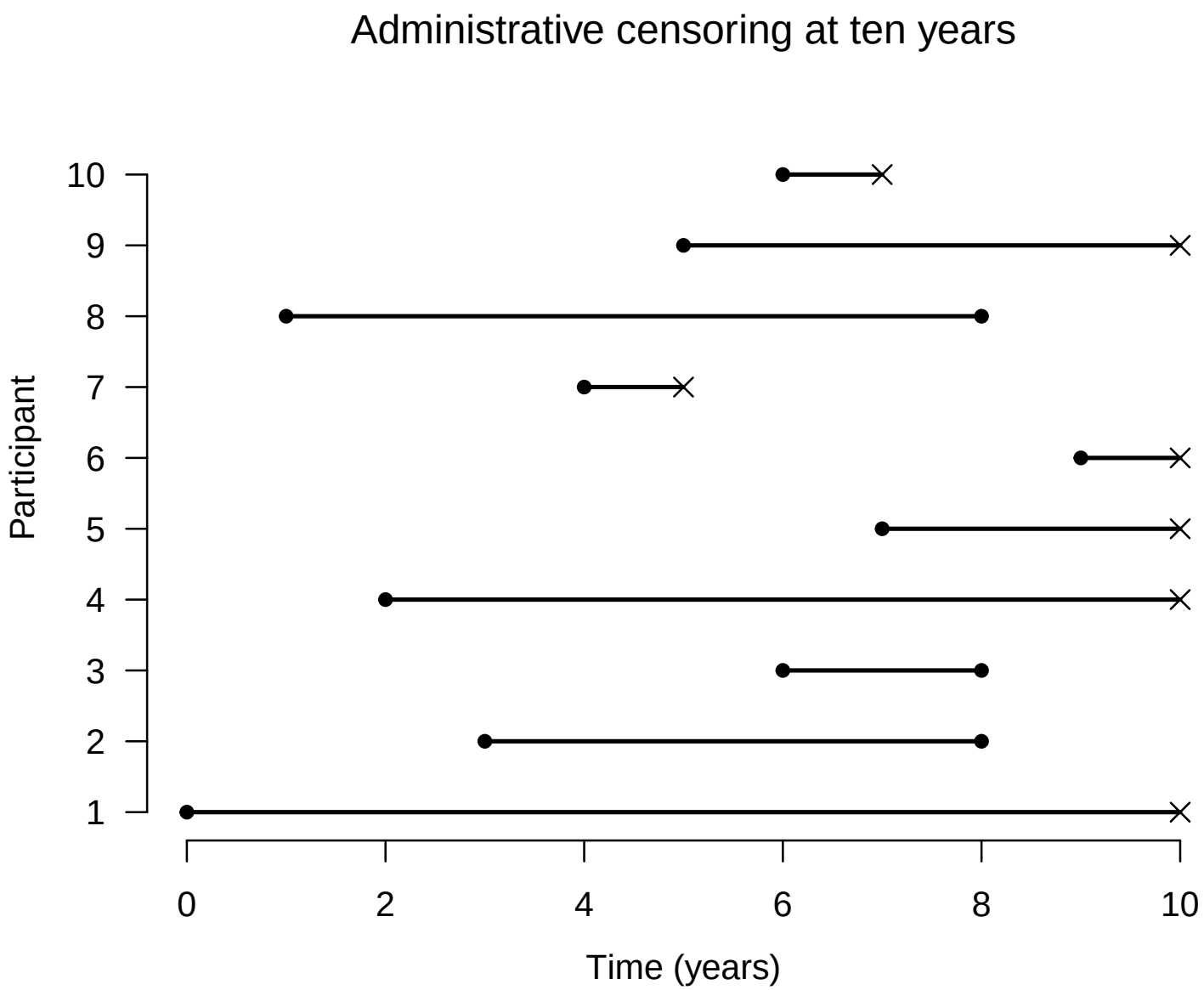
Version: 04/26/2025

# Table of Contents

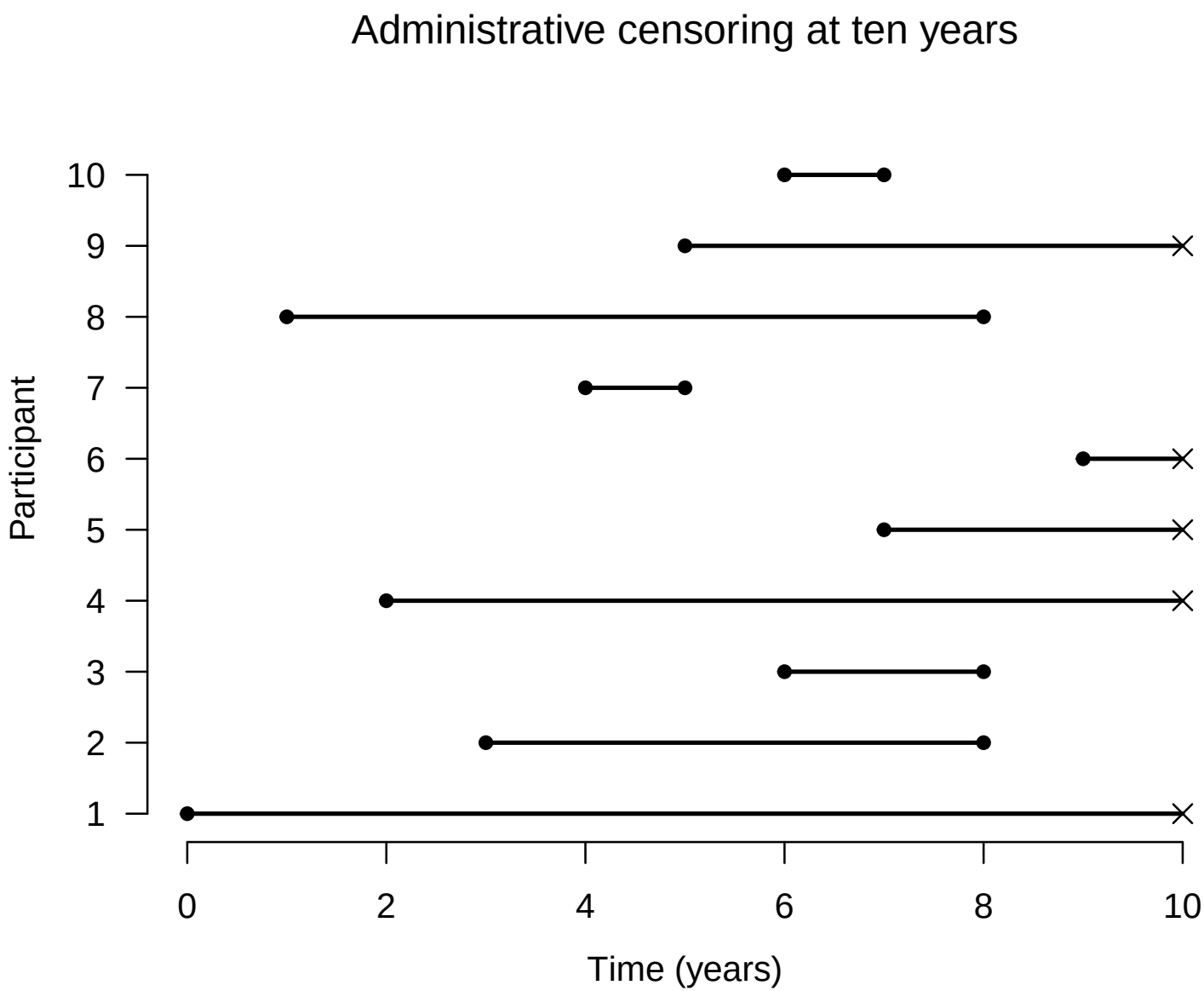
# Censoring

**Right-censoring**: A partially missing variable

- This is a special type of missing data in which it is only known that a variable exceeds a certain value (very common form of missingness when analyzing survival data).
  - The phrases "survival" and "failure time" analysis can be misnomers. Truly, it's time-to-event data that we're analyzing; the event can be either bad (death) or good (promotion).
- If we enroll, say, $N = 1000$ subjects to assess time-to-death when comparing two interventions, often end the study *before* all thousand participants have died.
- If a participant reaches the end of the study and is still alive, we stop following them.
- In these cases, we only know that they survived for *at least* a certain amount of time.


Administrative censoring at ten years
Participant
10
9
8
7
6
5
4
3
2
1
0
2
4
6
8
10
Time (years)

# Censoring


Administrative censoring at ten years
Participant
10
9
8
7
6
5
4
3
2
1
0
2
4
6
8
10
Time (years)

# Censoring

**Right-censoring**: A partially missing variable

- Time-to-event represented by two variables:
  - Time to event or censoring.
  - Indicator of event (binary).
- True times to event are given by $(T_1^0, T_2^0, \ldots, T_n^0)$.
- True times to censoring are given by $(C_1^0, C_2^0, \ldots, C_n^0)$.
- For each subject $i$, only see *either* $T_i^0$ *or* $C_i^0$—*never* both.
- Observed time to event/censoring: $T_i = \min\{T_i^0, C_i^0\}$.
- In a data set with time-to-event variables, you will often see variables for $T_i$ and an indicator for whether $T_i$ corresponds to an event (typically 1) or censoring (typically 0).

**Descriptive statistics**: Careful!

- We may be scientifically interested in the mean, median, standard deviation, etc. of time-to-event data.
- Descriptive statistics as we have been applying them are *not appropriate* for application to right-censored data.
  - Sample means, median, standard deviation, etc. do not account for the censoring properly.
- You need to tell R that you're working with right-censored data before trying to do anything with it. We'll do examples in a bit!

# Table of Contents

**Discrete hazards**:

- Cumulative failure rate over a window of time among those subjects who did not fail at the beginning of that window:

$$H(t, \Delta t) = \mathrm{P}(t \leq T^0 \leq t + \Delta t | t < T^0).$$

- Takes a little bit of work to get the idea:
  - Riddle 1: What is the probability of surviving to age 65?
  - Riddle 2: I'm 64 years old. What is the probability that I die within the next year?
  - Riddle 3: My risk of dying in the next year is about 50%. How old am I?
- . . . Once you get the idea, it becomes clear that the hazard is really quite a sensible metric.

**Working with right-censored data:** Hazards

- Ordered distinct observation times: $0 < t_1 < t_2 < \cdots < t_k$.
- $k$ intervals, $\Delta_j = (t_{j-1}, t_j]$. Estimated hazard in interval $j$:

$$\widehat{H}(t_j, \Delta_j) \quad = \quad \frac{D_j}{N_j} = \frac{\#\text{ events during interval}}{\#\text{ at risk at beginning of interval}}.$$

- Estimated probability of survival in interval, *given* survival until that interval:

$$\widehat{\mathrm{P}}(T^0 > t_j | T^0 > t_{j-1}) = 1 - \widehat{H}(t_j, \Delta_j).$$

**Working with right-censored data:** Hazards

- Cumulative probability of survival is a function of time.
- At a given time, $t$, survival rate is $S(t) = \mathrm{P}(T^0 > t)$.
- To estimate $S(t)$, we take the products of one-minus the estimated hazards in each interval, up until point $t$:

$$\begin{aligned}\widehat{S}(t_j) &= \prod_{k=1}^{j} \widehat{\mathrm{P}}(T^0 > t_k | T^0 > t_{k-1}) = \prod_{k=1}^{j} (1 - \widehat{H}(t_j, \Delta_j)) \\ &= \left(1 - \frac{D_j}{N_j}\right) \times \left(1 - \frac{D_{j-1}}{N_{j-1}}\right) \times \cdots \times \left(1 - \frac{D_1}{N_1}\right).\end{aligned}$$

- This is called the Kaplan-Meier estimator.

**By-hand example**:

- Not so bad with a small data set!

| ID | Obs. Time | Death? |
|---|---|---|
| 1 | 130 | 1 |
| 2 | 149 | 1 |
| 3 | 275 | 0 |
| 4 | 286 | 1 |
| 5 | 310 | 0 |
| 6 | 319 | 1 |
| 7 | 347 | 1 |
| 8 | 365 | 0 |

**By-hand example**: Not so bad with a small data set!

- Step 1: Organize in ascending order of observation time.

| Time | At risk | Events |
|---|---|---|
| 0 | 8 | 0 |
| 130 | 8 | 1 |
| 149 | 7 | 1 |
| 275 | 6 | 0 |
| 286 | 5 | 1 |
| 310 | 4 | 0 |
| 319 | 3 | 1 |
| 347 | 2 | 1 |
| 365 | 1 | 0 |

**By-hand example**: Not so bad with a small data set!

- Step 2: Include column for one minus hazard.

| Time | At risk | Events | 1 − Hazard |
|---|---|---|---|
| 0 | 8 | 0 | 1.0000 |
| 130 | 8 | 1 | 0.8750 |
| 149 | 7 | 1 | 0.8571 |
| 275 | 6 | 0 | 1.0000 |
| 286 | 5 | 1 | 0.8000 |
| 310 | 4 | 0 | 1.0000 |
| 319 | 3 | 1 | 0.6667 |
| 347 | 2 | 1 | 0.5000 |
| 365 | 1 | 0 | 1.0000 |

**By-hand example**: Not so bad with a small data set!

- Step 3: Include column for cumulative survival probability.

| Time | At risk | Events | $1 -$ Hazard | $\widehat{S}(t)$ |
|---|---|---|---|---|
| 0 | 8 | 0 | 1.0000 | 1.000 |
| 130 | 8 | 1 | 0.8750 | 0.8750 |
| 149 | 7 | 1 | 0.8571 | 0.7500 |
| 275 | 6 | 0 | 1.0000 | 0.7500 |
| 286 | 5 | 1 | 0.8000 | 0.6000 |
| 310 | 4 | 0 | 1.0000 | 0.6000 |
| 319 | 3 | 1 | 0.6667 | 0.4000 |
| 347 | 2 | 1 | 0.5000 | 0.2000 |
| 365 | 1 | 0 | 1.0000 | 0.2000 |

**By-hand example**: Not so bad with a small data set!

- Step 4: Remove the redundant rows (keep final row to know range of observation times).

| Time | At risk | Events | $1 -$ Hazard | $\widehat{S}(t)$ |
|---|---|---|---|---|
| 0 | 8 | 0 | 1.0000 | 1.000 |
| 130 | 8 | 1 | 0.8750 | 0.8750 |
| 149 | 7 | 1 | 0.8571 | 0.7500 |
| 286 | 5 | 1 | 0.8000 | 0.6000 |
| 319 | 3 | 1 | 0.6667 | 0.4000 |
| 347 | 2 | 1 | 0.5000 | 0.2000 |
| 365 | 1 | 0 | 1.0000 | 0.2000 |

- This is the basic method implemented by modern software. It's really that simple! No magic.

**Example**: MRI

- Government sponsored cohort study of adults aged 65 years and older.
- Many cardiovascular/cerebrovascular variables measured.
- `obstime`: time to death or censoring.
- `death`: indicator that participant died while on study (if $= 1$, then `obstime` is time until death; if $= 0$, then `obstime` is time until censoring).

## **R**: Plotting the Kaplan-Meier

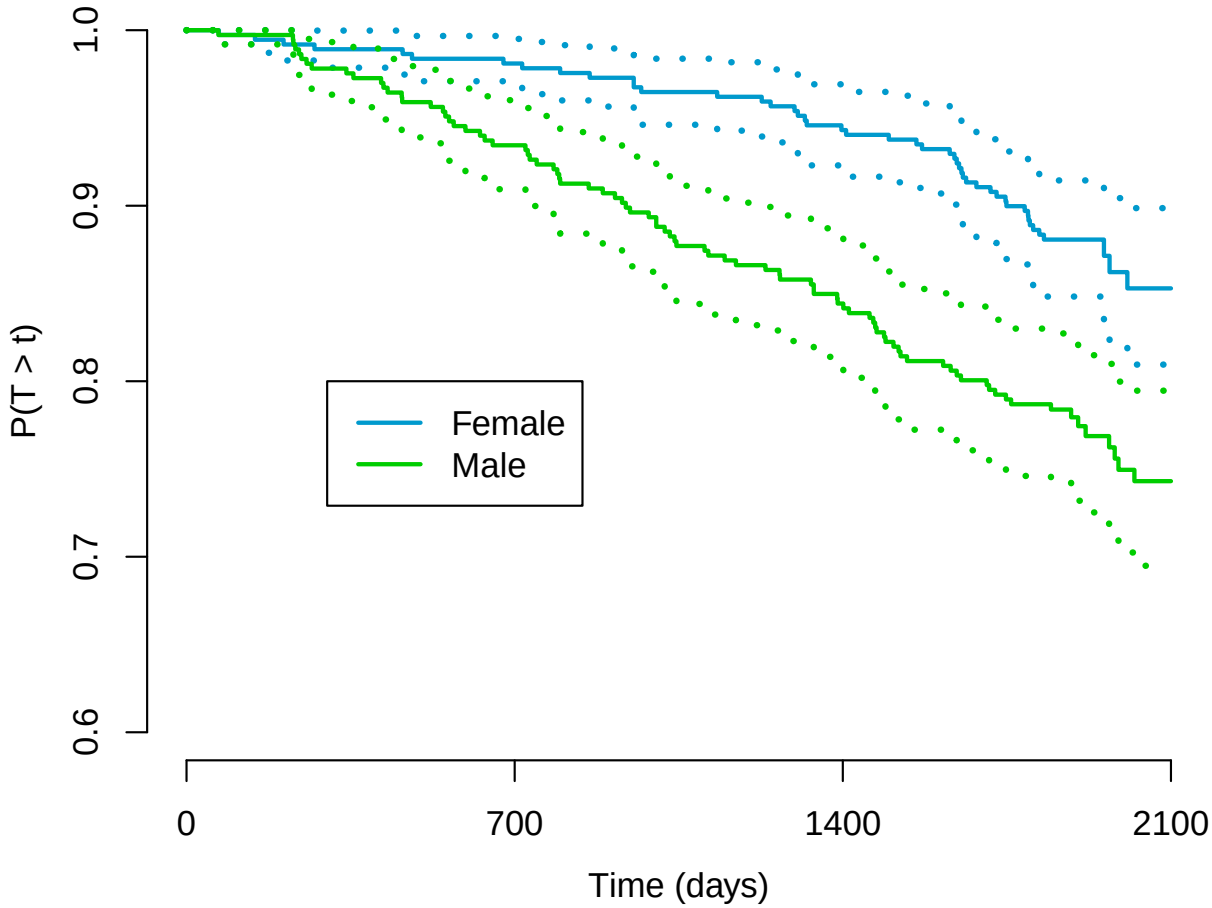
```
## Important library
library("survival")

## Read in data set
dat <- read.csv("mri.csv")

## Create a "survfit" object
km_fit <- survfit(Surv(obstime, death) ~ male, data=dat)
## Do "~ 1" instead if you just want to see the curve overall

## Plot Kaplan-Meier curves
plot(km_fit, conf.int = TRUE, col = c("deepskyblue3", "green3"),
     lty = c(1,3,3,  1,3,3), lwd = c(2,1,1,  2,1,1),
     xlab = "Time (days)", ylab = "P(T > t)",
     xlim = c(0,2100), ylim = c(0.6,1),
     xaxt = 'n', frame.plot = FALSE,)
axis(1, c(0,700,1400,2100))
```

**Example**: Kaplan-Meier by sex (MRI data)



**Example**: Kaplan-Meier by sex (MRI data)

- Note: the curves "stop" when there is no more data to support estimation.
- This is a nonparametric estimation method, meaning that it does not make any assumptions that the distribution of survival times belongs to some class of models indexed by a finite-dimensional parameter.
- Many descriptive statistics can be interpreted graphically.
  - Proportion surviving until $t$.
  - Proportion surviving beyond $t$.
  - Percentile of the survival distribution.
- Restricted mean survival time through time $t$ (area under the curve until time $t$).

## **R**: Proportions

```
## Detail on KM curves with particular times called out
KM <- summary(km_fit, times = c(100,1500,2100,2159,2170))

> KM
Call: survfit(formula = Surv(obstime, death) ~ male, data = dat)

                male=0
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
  100    368       1    0.997 0.00271        0.992        1.000
 1500    346      22    0.938 0.01259        0.913        0.963
 2100     32      24    0.853 0.02273        0.809        0.899

                male=1
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
  100    365       1    0.997 0.00273        0.992        1.000
 1500    301      64    0.822 0.01998        0.784        0.863
 2100     49      21    0.743 0.02543        0.695        0.795
 2159      1       0    0.743 0.02543        0.695        0.795

## Notice that it ignored my request to include Day 2159 for male = 0.
## ...why?
```

## **R**: Quantiles

```
## Estimating stratum-specific quantiles of the survival distribution
> quantile(km_fit, c(0.1, 0.25, 0.4))
$quantile
         10   25 40
male=0 1749   NA NA
male=1  937 1988 NA

$lower
         10   25 40
male=0 1643   NA NA
male=1  732 1748 NA

$upper
         10 25 40
male=0 2007 NA NA
male=1 1235 NA NA

## Once again, some of my requests were rejected.
## Is R just deciding to be rude?
```

## **R**: Restricted mean survival time

```
## Restricted mean survival time to 500 days (abridged output)
> print(km_fit, print.rmean=TRUE, rmean=500)

         n events rmean* se(rmean) median 0.95LCL 0.95UCL
male=0 369     47    496      1.81     NA      NA      NA
male=1 366     86    492      2.28     NA      NA      NA
    * restricted mean with upper limit =  500

## Restricted mean survival time to 1000 days (abridged output)
> print(km_fit, print.rmean=TRUE, rmean=1000)

         n events rmean* se(rmean) median 0.95LCL 0.95UCL
male=0 369     47    985      5.06     NA      NA      NA
male=1 366     86    955      8.04     NA      NA      NA
    * restricted mean with upper limit =  1000

## Restricted mean survival time to 2159 days (abridged output)
> print(km_fit, print.rmean=TRUE, rmean=2159)

         n events rmean* se(rmean) median 0.95LCL 0.95UCL
male=0 369     47   2050      17.6     NA      NA      NA
male=1 366     86   1899      27.6     NA      NA      NA
    * restricted mean with upper limit =  2159
```

**Note**: R *may* allow you to go beyond the maximum observation time with an exponential extrapolation (not advisable).

**Key ideas and assumptions**:

- That the (continuous) survival function, and descriptive statistics off of which it is based, can be estimated with a discrete function is *not* immediately evident.
  - Requires rigorous empirical process/martingale theory (not the focus of this course).
- Key assumption: *Non-informative censoring.*
  - Participants available at the start of each time interval are a random sample of the population surviving to that time (censored subjects neither more nor less likely to have an event in the immediate future).

**Examples**: Potentially informative censoring

- Subjects in RCT are withdrawn due to treatment failure.
  - Likely would die sooner than those remaining.
- Subjects in RCT in a fatal condition are lost to follow up when they go on vacation.
  - Likely they are healthier than those remaining.
- Leukemia patients in RCT of bone marrow transplantation censored if they die of infections rather than cancer.
  - May have had more effective regimen to wipe out cancer.

**Key ideas**:

- Time zero must be defined. Often consider time at risk.
- On what scale?
  - Study time: time since diagnosis or treatment (e.g., RCT).
  - Age: time since birth (epidemiologic).
  - Study time: time since first exposure (epidemiologic).

# Table of Contents

# The log-rank test

**Comparing survival distributions**: Two groups

- Let $N_{kj}$ and $O_{kj}$ denote the number of subjects in group $j$ at risk at the start of interval $k$, and the number of observed events in the group $j$ during interval $k$, respectively.
- In turn, let $N_j = \sum_k N_{kj}$ and $O_j = \sum_k O_{kj}$.
- Null hypothesis: $H_0 : \lambda(t|X = 0)/\lambda(t|X = 1) = 1$.
- Under $H_0$, $O_{kj}$ is hypergeometric with mean $E_{kj} = N_{kj}O_j/N_j$ and variance $V_{kj} = E_{kj}(N_j - O_j)(N_j - N_{kj})/(N_j(N_j - 1))$.
- The log-rank statistic:

$$Z = \frac{\sum_j (O_{ij} - E_{ij})}{\sqrt{\sum_j V_{ij}}} \xrightarrow{d} \mathcal{N}(0,1) \qquad (Z^2 \dot{\sim} \chi^2_1).$$

- This is a **non-parametric** test of (in)equality of the two survival distributions.

**Comparing survival distributions**: Two groups

```
## Log-rank test
> survdiff(Surv(obstime, death) ~ male, data=dat)

Call:
survdiff(formula = Surv(obstime, death) ~ male, data = dat)

         N Observed Expected (O-E)^2/E (O-E)^2/V
male=0 369       47     68.8      6.89      14.3
male=1 366       86     64.2      7.38      14.3

 Chisq= 14.3  on 1 degrees of freedom, p= 2e-04
```

# Table of Contents

**Discrete hazards**:

- Cumulative failure rate over a window of time among those subjects who did not fail at the beginning of that window:

$$H(t, \Delta t) = \mathrm{P}(t \leq T^0 \leq t + \Delta t | t < T^0).$$

  - Car speed analogy: 15 miles over a 20 minute period.

- Average failure rate (per unit of time) in a window of time among those subjects who did not fail at the beginning of that window:

$$h(t, \Delta t) = \frac{\mathrm{P}(t \leq T^0 \leq t + \Delta t | t < T^0)}{\Delta t}.$$

  - Car speed analogy: Average 45 miles per hour over a 20 minute period.

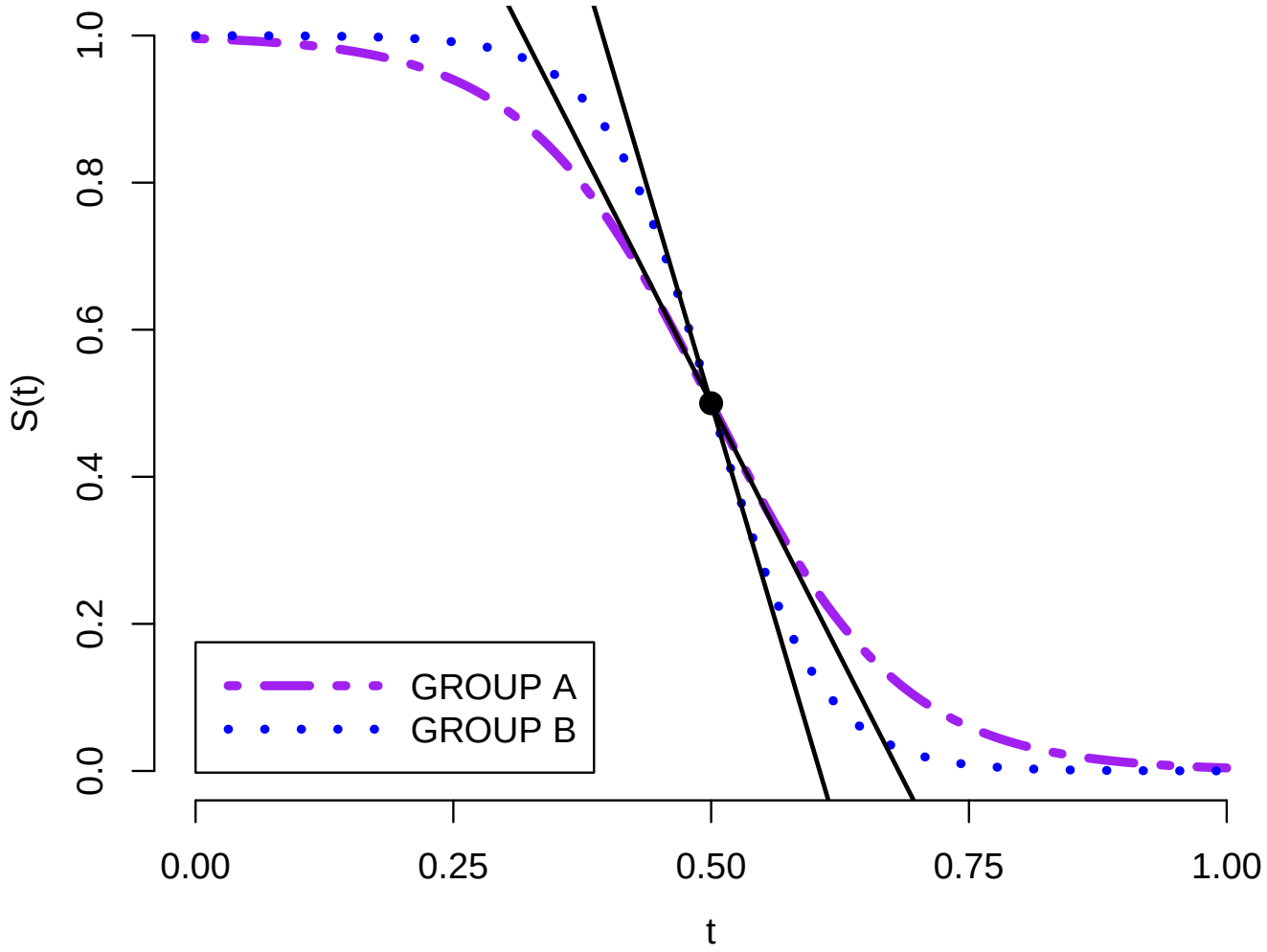
**Hazard rate**:

- Continuous hazard rate (instantaneous failure rate):

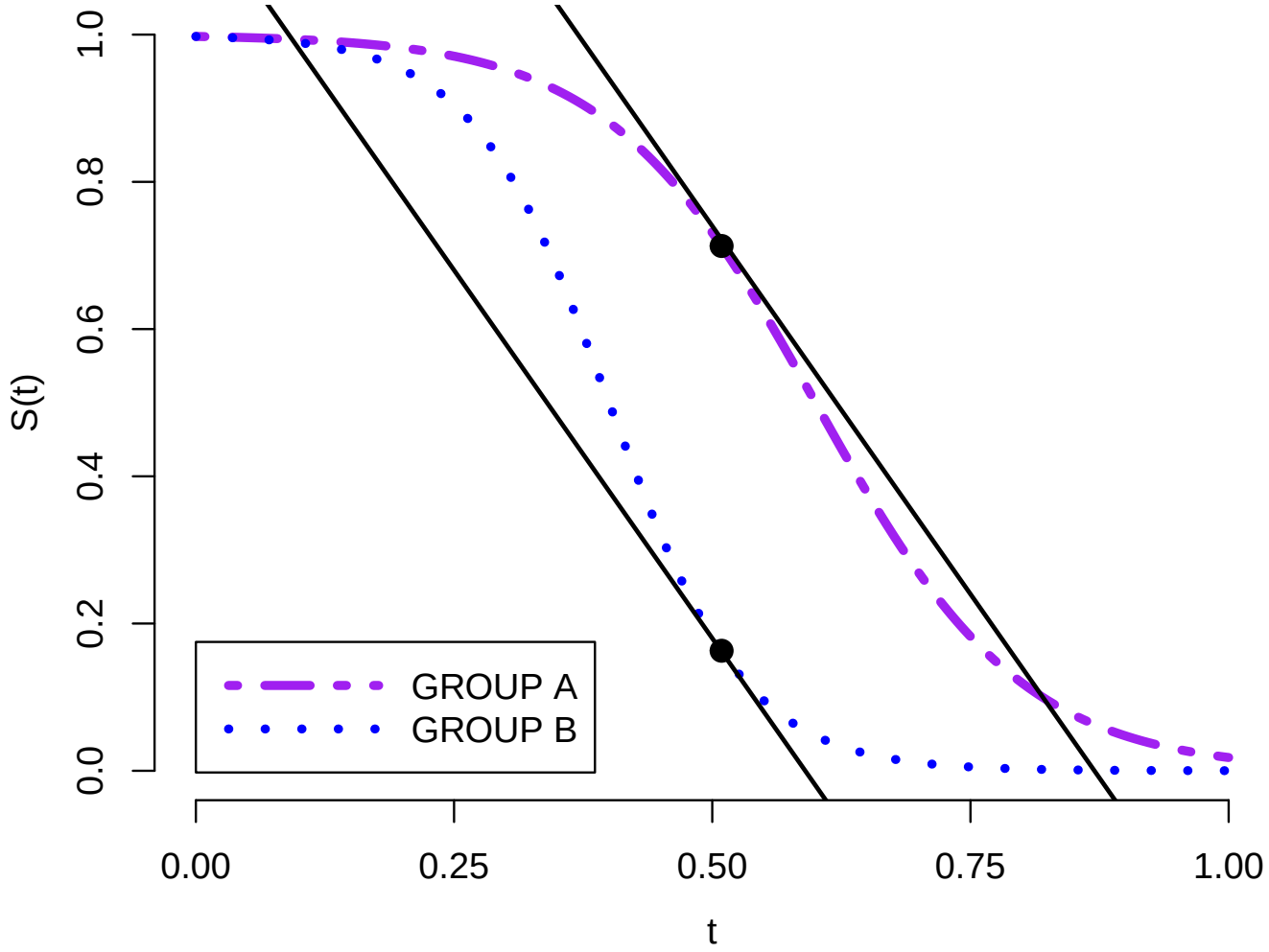
$$\lambda(t) = \lim_{\Delta t \longrightarrow 0} \frac{\mathrm{P}(t \leq T^0 \leq t + \Delta t | t < T^0)}{\Delta t} = \frac{f(t)}{S(t)}.$$

  - Car speed analogy: Going 20 miles per hour at $t = 1\text{m}30\text{s}$.
- Graphically: negative slope of survival divided by height.

**Question**: Which group has the higher hazard at time $t = 0.5$?



**Question**: Which group has the higher hazard at time $t = 0.5091$?



**Notation and relationships**:

- Survival function: $S(t) = P(T > t)$.
- Hazard rate:

$$\begin{aligned} \lambda(t) &= \lim_{\Delta t \longrightarrow 0} \frac{\mathrm{P}(t \leq T^0 \leq t + \Delta t | t < T^0)}{\Delta t} \\ &= -\frac{\partial \log(S(t))}{\partial t} = -\frac{\partial S(t)/\partial t}{S(t)} \end{aligned}$$

- Cumulative hazard: $\Lambda(t) = \int_0^t \lambda(u)du = -\log(S(t))$.
  - Equivalently, $S(t) = \exp(-\Lambda(t))$.

# Proportional hazards regression

**Comparing hazards**: Single binary predictor, $X$

- Two groups ($X = 0$; $X = 1$).
- Baseline hazard: $\lambda(t|X = 0) = \lambda_0(t)$ (an entire function).
- Model: $\log(\lambda(t|X = x)) = \log(\lambda_0(t)) + \beta x$.
- Assumes parallel log-hazards:

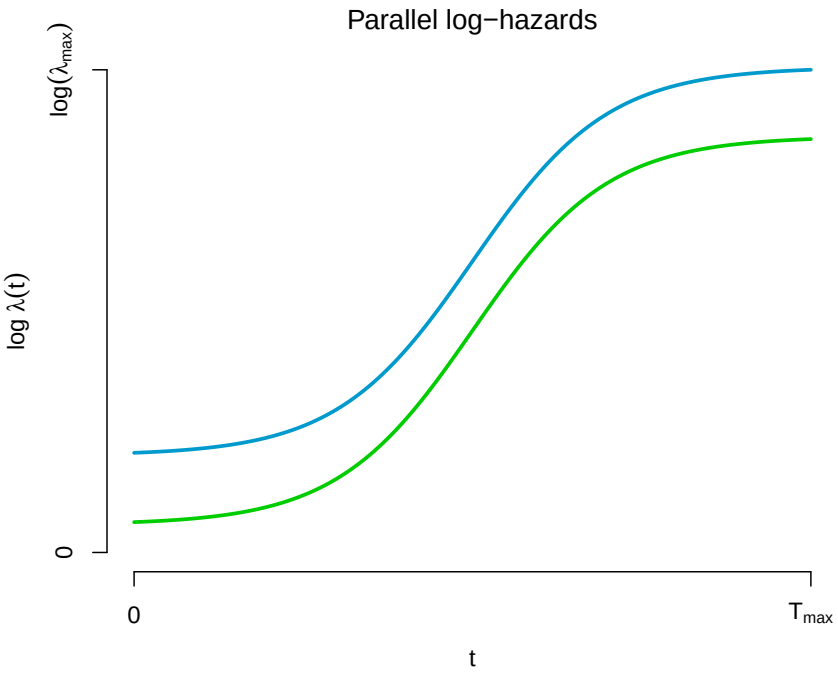
$$\log\left(\lambda(t|X = 1)\right) = \log(\lambda_0(t)) + \beta$$

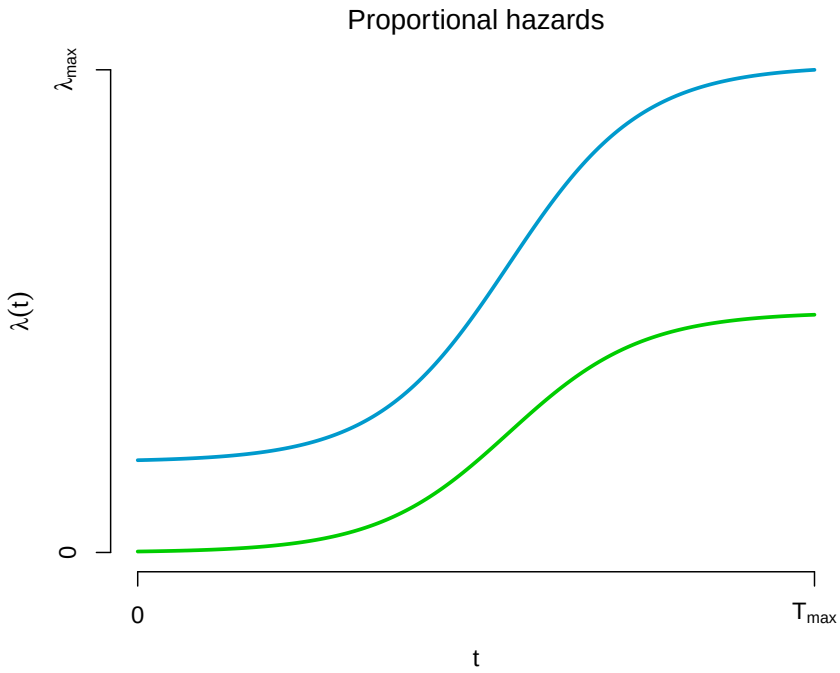
  - More commonly expressed as *proportional hazards*:

$$\lambda(t|X = 1) \quad = \quad \lambda_0(t)e^{\beta} \quad \Longleftrightarrow \quad \frac{\lambda(t|X = 1)}{\lambda(t|X = 0)} = e^{\beta}$$

- $e^{\beta}$: hazard ratio comparing strata defined by $X$.
- Not a saturated model (even though only two groups).

# Proportional hazards regression



# Proportional hazards regression

**Comparing hazards**: Single exposure, $X$

- Model: $\log(\lambda(t|X=x) = \log(\lambda_0(t)) + \beta x$.
- Equivalently, proportional hazards:

$$\lambda(t|X=x) \quad = \quad \lambda_0(t)e^{\beta x} \quad \Longleftrightarrow \quad \frac{\lambda(t|X=x+1)}{\lambda(t|X=x)} = e^{\beta}$$

- $e^{\beta}$: hazard ratio comparing strata differing in their value of $X$ by one unit.

**Comparing hazards**: Single continuous exposure, $X$

- Cox proportional hazards model:

$$\lambda(t|X = x) = \lambda(t|X = 0)\exp(\beta x) = \lambda_0(t)\exp(\beta x).$$

- Likelihood of event to be observed at time $T_i$:

$$L_i(\beta; X_i) = \frac{\lambda(T_i|X_i)}{\sum_{j:T_j \geq T_i} \lambda(T_i|X_j)} = \frac{\exp(X_i\beta)}{\sum_{j:T_j \geq T_i} \exp(X_j\beta)}.$$

- (Partial) likelihood:

$$\mathcal{L}(\beta; \mathbf{X}) = \prod_{i:\text{uncensored}} L_i(\beta; X_i).$$

# Proportional hazards regression

**Comparing hazards**: MRI example (gender)

- Single predictor:
  - $X = 0$: females; $X = 1$: males.
- Cox model: $\log(\lambda(t|X = 1)) = \log(\lambda_0(t|X = 0)) + \beta$.
- R will perform proportional hazards regression: `coxph()`.
  - `sandwich()`: compatible.
  - My custom `LinCom()` function is therefore compatible.

## **R**: Cox model

```
## Fit model (abridged output)
model <- coxph(Surv(obstime, death) ~ male, data = dat)

> summary(model)

Call:
coxph(formula = Surv(obstime, death) ~ male, data = dat)

  n= 735, number of events= 133

      coef exp(coef) se(coef)    z Pr(>|z|)
male 0.674     1.962    0.182 3.71  0.00021 ***
---

     exp(coef) exp(-coef) lower .95 upper .95
male      1.96       0.51      1.37       2.8

Concordance= 0.586  (se = 0.021 )
Likelihood ratio test= 14.4  on 1 df,   p=1e-04
Wald test            = 13.8  on 1 df,   p=2e-04
Score (logrank) test = 14.3  on 1 df,   p=2e-04
```

**Comparing hazards**: Multiple predictors

- Cox proportional hazards model:

$$\lambda(t|X_1 = x_1, \cdots, X_K = x_K) = \lambda_0(t)\exp(\beta_1 x_1 + \cdots + \beta_K x_K).$$

- Likelihood of event to be observed at time $T_i$:

$$L_i(\boldsymbol{\beta}; \mathbf{X}_i) = \frac{\lambda(T_i|X_i)}{\sum_{j:T_j \geq T_i} \lambda(T_i|X_j)} = \frac{\exp(\sum_{k=1}^K X_{ik}\beta_k)}{\sum_{j:T_j \geq T_i} \exp(\sum_{k=1}^K X_{jk}\beta_k)}.$$

- (Partial) likelihood:

$$\mathcal{L}(\boldsymbol{\beta}; \mathbf{X}) = \prod_{i:\text{uncensored}} L_i(\boldsymbol{\beta}; \mathbf{X}_i).$$

# Proportional hazards regression

**Cox proportional hazards model**: Ideas and assumptions

- Assumes any right-censoring is non-informative (conditional on covariates in the model).
- Assumes proportional hazards.
  - Able to bypass estimation of $\lambda_0(t)$ (baseline hazard).
  - The Cox model is said to be *semiparametric*.
  - Baseline hazard serves something like an intercept, except it is an entire function.
- Cox model borrows information over time (relative hazard of event presumed time-invariant), and across covariates.

**Example**: MRI

- This study provides evidence that males tend to have a higher hazard of death ($p < 0.001$). We estimate that the males have an 96.2% higher hazard of death as compared to the females. Based on a 95% CI, this estimate would not be surprising if in truth the hazard were anywhere between 37.8% higher and 179% higher in males.
- This study provides evidence that males tend to have a higher hazard of death ($p < 0.001$). We estimate the hazard ratio to be 1.962, with the males having the higher estimated hazard. Based on a 95% CI, this estimate would not be judged unusual if the true hazard ratio were between 1.378 and 2.79.

**Example**: MRI

- Suppose we want to build a predictive model for hazard of death and include the following covariates:
  - Age
  - Gender
  - General (self) view of health (categorical)
  - Smoking history (pack years)
  - Diabetes status
  - Alcohol consumption
  - Physical activity

## **R**: Cox model

```
## General health as a factor variable
dat$genhlth <- factor(dat$genhlth)

## Fit model (abridged output)
> model <- coxph(Surv(obstime, death) ~ age + male + genhlth + packyrs + diabetes +
                 alcoh + physact, data = dat)
> summary(model)
Call:
coxph(formula = Surv(obstime, death) ~ age + male + genhlth +
    packyrs + diabetes + alcoh + physact, data = dat)

  n= 734, number of events= 132
   (1 observation deleted due to missingness)


          exp(coef) exp(-coef) lower .95 upper .95
age           1.075      0.930     1.046      1.10
male          1.833      0.546     1.264      2.66
genhlth2      1.015      0.986     0.522      1.97
genhlth3      1.105      0.905     0.584      2.09
genhlth4      2.310      0.433     1.176      4.54
genhlth5      2.635      0.380     0.919      7.56
packyrs       1.010      0.990     1.005      1.02
diabetes      1.695      0.590     1.091      2.63
alcoh         0.956      1.046     0.914      1.00
physact       0.973      1.028     0.887      1.07
```

**Example**: MRI (example interpretation)

- Comparing strata differing in smoking history by one pack year, but of the same age, gender, self-view of health, diabetes status, alcohol consumption, and physical activity, the group with a more extensive smoking history has an estimated 1.00% higher hazard of death. Based on a 95% confidence interval, this estimate would not be judged unusual if in truth the hazard for the stratum with a more extensive smoking history were between 0.448% and 1.55% higher. This study provides evidence of a (covariate-adjusted) association between smoking history and hazard of death ($p < 0.001$).

**R**: Joint test for Cox model

```
## Sandwich library
library("sandwich")

## Custom JointTest() function from prior notes
> JointTest(idxtest = c(3:6),
            coefs = coef(model),
            varmat = sandwich(model))
chi2.stat         p
16.224000  0.002733
```

**Further notes and advice**:

- Hazard ratios are not collapsible.
  - Better to adjust for covariates suspected to be associated with hazard, but not necessarily for reasons of precision.
- I am expecting that you are able to generalize some of the basic ideas regarding log-transformations of predictors, interaction terms, etc. with ease.
  - For instance, an exponentiated interaction term may be interpreted as a ratio of hazard ratios, etc.

# Table of Contents

# Assessing proportional hazards

**Proportional hazards**: Parallel log-negative-log-survival

- Under the proportional hazards assumption:

$$\begin{aligned}
\lambda_1(t) &= \exp(\beta)\lambda_0(t) \\
\Lambda_1(t) &= \exp(\beta)\Lambda_0(t) \\
\exp(-\Lambda_1(t)) &= \exp(-\Lambda_0(t)\exp(\beta)) = [\exp(-\Lambda_0(t))]^{\exp(\beta)} \\
S_1(t) &= [S_0(t)]^{\exp(\beta)} \\
-\log[S_1(t)]) &= -\exp(\beta)\log[S_0(t)] \\
\log(-\log[S_1(t)]) &= \log(-\log[S_0(t)]) + \beta
\end{aligned}$$

- Note: $\log(-\log(x))$: Complementary log-log link.

# Assessing proportional hazards

**Note**:

- We learn from this that if the proportional hazards assumption holds, then $\log(-\log(S(t|X=x))$ should be parallel between strata of $X$ over time, $t$.
- Often consider $-\log(-\log(S(t|X=x))$ across $\log(t)$ for convenience.
- In R, you can extract the estimates from the Kaplan-Meier curve.

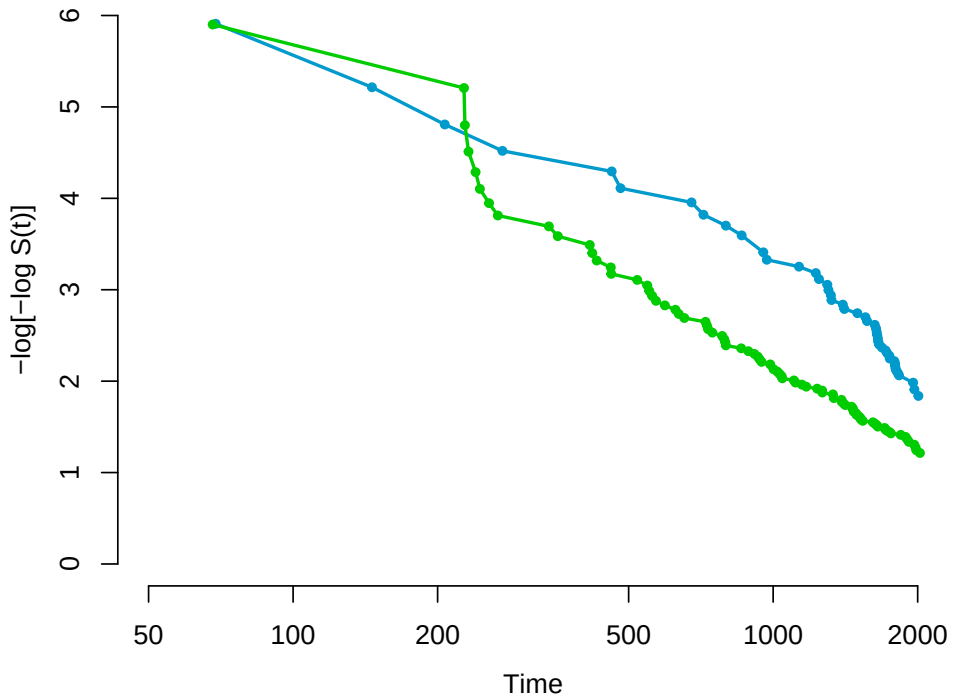
## **R**: Graphical assessment of proportional hazards

```
## Pull out full summary of Kaplan-Meier
KMSummary <- summary(km_fit)

## Create plot
plot(KMSummary$time[1:46],-log(-log(KMSummary$surv[1:46])), log = 'x',
     xlim = c(50,2000), ylim = c(0,6), type = "o",
     col = "deepskyblue3", lwd = 2, cex = 0.8, pch = 20,
     frame.plot = FALSE, xlab = "Time",
     ylab = "-log[-log S(t)]",
     main = "")
lines(KMSummary$time[47:130],-log(-log(KMSummary$surv[47:130])),
      col = "green3", lwd = 2, cex = 0.8, pch = 20,
      type = "o")
```

# Assessing proportional hazards

**Example**: MRI (proportional hazards)




**Note**: The curves come closer together over time providing evidence of a potential violation to proportional hazards. Not emphasizing left side—very few events for $t < 500$.

# Assessing proportional hazards

**Schoenfeld residuals**:

- After running a Cox model, one can test the proportional hazards assumption by variable and/or overall: `cox.zph(model)`.
- Uses *Schoenfeld residuals* (choosing not to elaborate on how these are defined, although you can look it up if you're interested).
- As with all hypothesis tests for assumptions, this suffers from the problem that you cannot prove the null hypothesis true, which is what you would want.
- Also, minute, irrelevant departures from assumptions will be detected with a sufficiently large sample.
- Don't take these sorts of hypothesis tests too seriously. Graphical evaluation is often adequate for giving you a sense of whether your conclusion is likely to be wildly off (particularly in simple settings).

## **R**: Assessing proportional hazards

```
## Test of Schoenfeld residuals
> cox.zph(model)

            chisq df      p
age      8.38e-08  1 0.9998
male     7.00e+00  1 0.0082
genhlth  2.83e+00  4 0.5866
packyrs  6.55e-03  1 0.9355
diabetes 8.39e-01  1 0.3596
alcoh    1.29e+00  1 0.2563
physact  2.62e-02  1 0.8713
GLOBAL   1.17e+01 10 0.3038

## We had seen graphical evidence for a violation
## by male = 0/male = 1.
## ...But was it an egregious violation?
```

# Assessing proportional hazards

**Note**:

- A sign of a pretty major violation to the proportional hazards assumption is if the Kaplan-Meier curves cross (and I don't mean that in the case where they have a lot of overlap).
  - Draw the picture!
- Can you come up with an example of where something like this might happen?

# Assessing proportional hazards

**Notes**:

- Incorrect line of reasoning: There is evidence of a departure from proportional hazards, so I can't claim an association.
  - The only way the proportional hazards assumption can be violated is if there is an association in the first place. Think of linearity as an example.
- We don't panic about evidence of minor violations. But we don't ignore evidence of extraordinary violations either.
- Don't let statistics force you into doing bad science.

# Summary

**This unit**:

- Censoring.
- Kaplan-Meier.
- Discrete and continuous hazards.
- Cox proportional hazards.
- Assumptions.
- There is a reason why survival analysis usually gets a full semester of treatment. There is *so much* we did not cover:
  - Time-dependent treatment.
  - Cumulative incidence.
  - Competing risks.
  - Accelerated failure time models.
  - Left- and interval-censoring.
  - Truncation.
- This unit is just to orient you to the basics.

# Summary

**So far**:

- Review.
- Simple linear regression.
- Multiple linear regression (foundations).
- Multiple linear regression (interactions and strata).
- Transformations and basis expansions.
- Regression with binary outcomes.
- Regression with nominal, ordinal, and count outcomes.
- Introduction to clustered data.
- Methods for time-to-event outcomes.

# Summary

**Coming up**:

- Predictive capacity of regression models.