Survival Analysis

Day 3

- Introduction to survival analysis
- Cox proportional hazards model
  - Parametric survival models

Breakout Session #3

1. What are two features of survival outcomes that require them to be analyzed with specific methods? What would be two consequences of applying standard regression methods?
2. When is the Kaplan-Meier estimate 0?
3. A clinical investigator wants to report the proportion of patients who experienced the event during the study. What is wrong with this estimate and what would you recommend reporting instead to describe the cohort's risk?
4. What are two benefits of using a parametric PH survival model compared to the Cox PH model?
5. What's your favourite animal?

Features of Survival Data

- The outcome of interest is the time until an event occurs
  - Time origin, time scale, definition of the event
- **Goal:** Estimate risk of the event
- Most important characteristic of time-to-event outcomes is **censoring**
- The event of interest is not observed for all subjects under study
- Implications:
  - Standard analysis techniques (sample average, t-test, linear regression) cannot be used
  - Inferences may be sensitive to misspecification of the distribution of the event times

Censoring

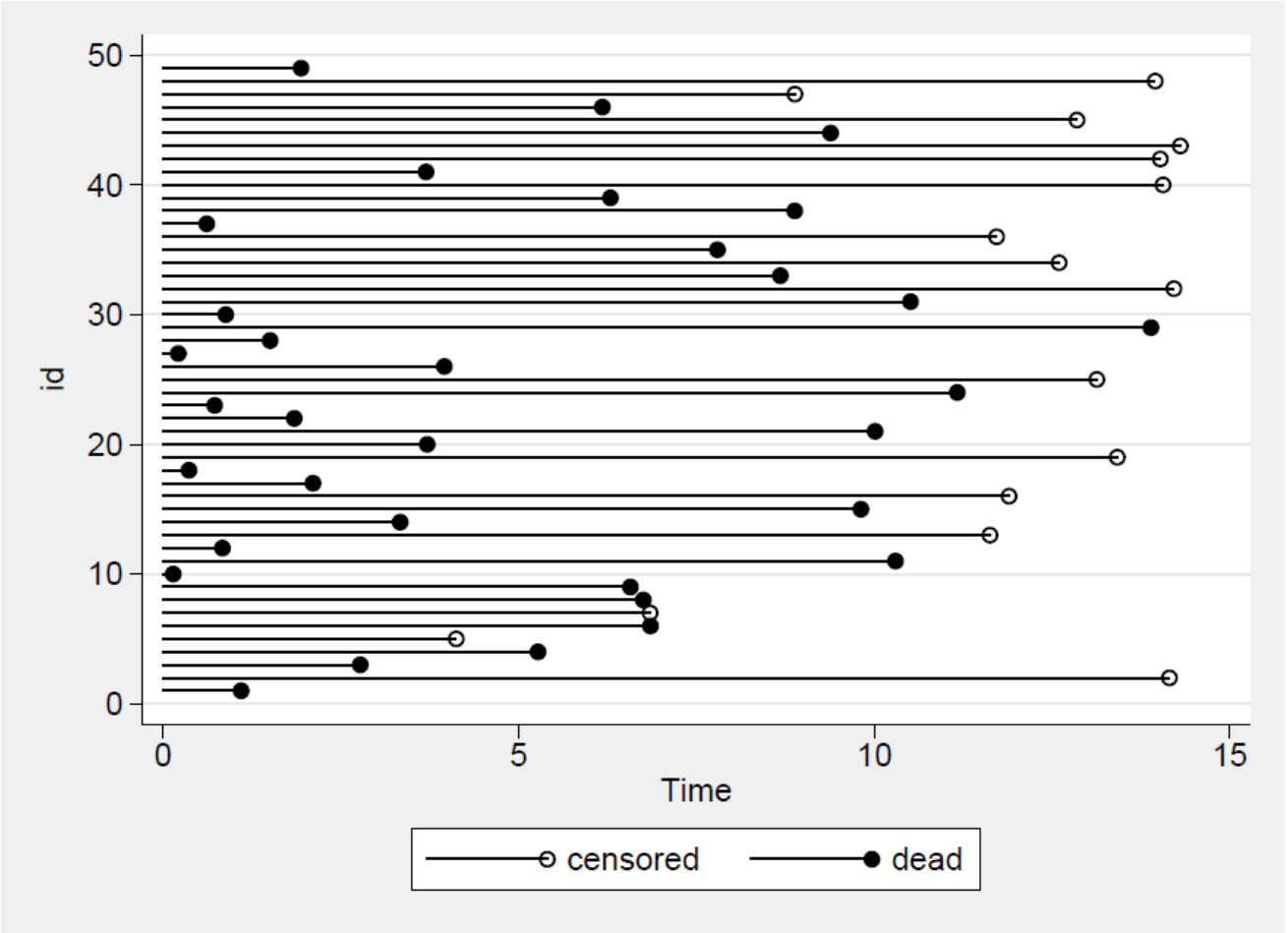
- Censoring can arise when:
  a) Patient has not yet experienced the outcome by the end of the study
  b) Patient is lost to follow-up during the study
  c) Patient experiences a different event that makes further follow-up impossible
- Observed censoring times underestimate the true survival times

Types of Censoring Mechanisms

1. Location of the true event time with respect to the censoring time: Right, left, interval

- **Right censoring:** For a subset of subjects, the event of interest is only known to occur **after** a certain point
- **Left censoring:** For a subset of subjects, the event of interest is only known to occur **before** a certain point
- **Interval censoring:** For a subset of subjects, the event of interest is only known to occur **between** two certain points

Right Censoring



Types of Censoring Mechanisms

2. Probabilistic relation between the true event time and the censoring time: informative and non-informative (similar to MNAR and MAR)

- **Informative censoring:** When a subject withdraws from the study for reasons directly related to their expected failure time
- **Non-informative censoring:** When a subject withdraws from the study for reasons not related to their prognosis but can depend on covariates

We will focus on non-informative right censoring!

Notation

- For subject i
  - T_i^* is the "true" event time
  - C_i is the censoring time (e.g., end of study or random censoring time)

- Data available for each subject:
  - Observed event time: $T_i = \min(T_i^*, C_i)$
  - Event indicator: $\delta_i = I(T_i^* \leq C_i)$
  - $\delta_i = 1$ if event; $\delta_i = 0$ if censored

- **Goal:** Make valid inferences for T_i^* but using only $\{T_i, \delta_i\}$

Basic functions in Survival Analysis

- Survival data is usually modelled using two related functions:
  - **Survival** $S(t)$: probability than a subject survives from baseline to a specified future time t
  - **Hazard** $h(t)$: the instantaneous risk of having an event at time t for an individual who has survived to that time
- Survival function focuses on not having the event
- Hazard function focuses on the event occurring

Survival Function

- **Survival function:** The probability of being alive up to time t

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

- Decreasing function of time
- When there is no censoring:

$$\hat{S}_n(t) = \frac{\#\text{ individuals alive at } t}{\text{total sample size}}$$

Kaplan-Meier Survival Estimate

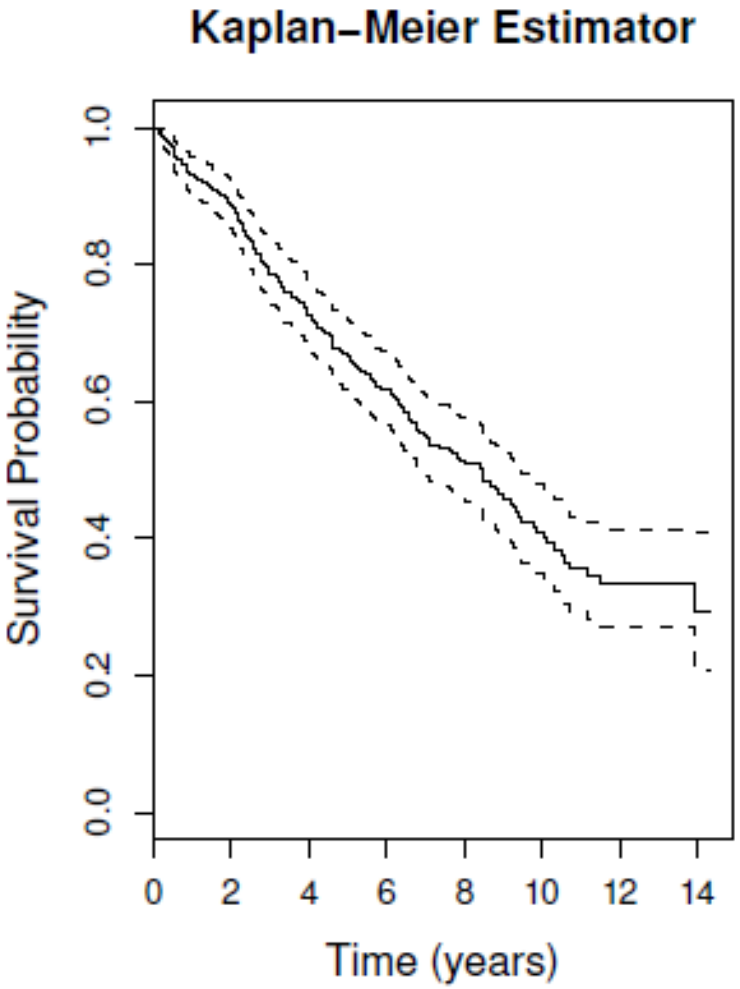
- **Kaplan-Meier** (or product-limit) method
- Non-parametrically estimates the survival probability using the observed survival times (censored and uncensored)

- Let $t_1, t_2, \ldots, t_k$ be the unique event times in the sample
- **Kaplan-Meier estimator**

$$\hat{S}_{KM}(t) = \prod_{i:t_i \leq t} \left(1 - \frac{d_i}{r_i}\right)$$

- r_i: # subjects still at risk at t_i
- d_i: # events at t_i

Kaplan-Meier Survival Estimate



Hazard Function

- **Hazard function:** The instantaneous risk of an event at time t, given that the event has not occurred until t

$$h(t) = \lim_{dt \to 0} \frac{\Pr(t \leq T^* < t + dt | T^* \geq t)}{dt}, \quad t > 0$$

- The hazard function is NOT a probability: $h(t) \in [0, \infty)$
- It can be interpreted as the expected number of events per individual per unit of time
- It is the event rate conditional on still being at risk
- Many survival models estimate (relative) differences in hazard rates

Relationship between the functions

- The survival function is connected to the hazard via

$$S(t) = \exp\left\{-\int_0^s h(s)ds\right\} \qquad h(t) = -\frac{d}{dt}[\log S(t)]$$

- Cumulative hazard function

$$H(t) = \int_0^t h(t)dt = -\log S(t) \qquad S(t) = \exp\{-H(t)\}$$

- Nelson-Aalen estimator

$$\hat{H}_{NA}(t) = \sum_{i:t_i \le t} \frac{d_i}{r_i};$$

Example: PBC data

- The primary package in R for the analysis of survival data is the **survival** package
- A key function that is used to specify the available event time information in a sample is Surv()
- For right censored failure times, we need to provide the observed follow-up times and the event indicator, which equals 1 for true failure times and 0 for right censored times

Surv([observed follow-up time], [event indicator])

- In our PBC example: Surv(years, status2)

Example: PBC data

- Kaplan-Meier estimate: survfit

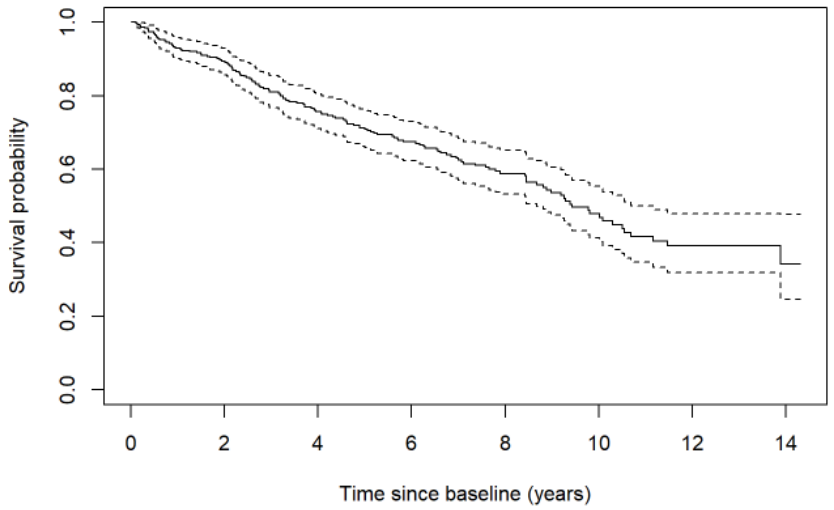
```
library(survival)
surv.pbc <- survfit(Surv(years, status2) ~ 1, data=dat.id)
summary(surv.pbc)
```

```
## Call: survfit(formula = Surv(years, status2) ~ 1, data = dat.id)
##
##    time n.risk n.event survival std.err lower 95% CI upper 95% CI
##   0.112    312       1    0.997 0.00320        0.991        1.000
##   0.140    311       1    0.994 0.00452        0.985        1.000
##   0.194    310       1    0.990 0.00552        0.980        1.000
##   0.211    309       1    0.987 0.00637        0.975        1.000
##   0.301    308       1    0.984 0.00711        0.970        0.998
##   0.361    307       1    0.981 0.00778        0.966        0.996
##   0.375    306       1    0.978 0.00838        0.961        0.994
##   0.383    305       1    0.974 0.00895        0.957        0.992
##   0.490    304       1    0.971 0.00948        0.953        0.990
```

r_i d_i

Example: PBC data

```
plot(surv.pbc, ylab="Survival probability", xlab="Time since baseline (years)")
```




```
summary(surv.pbc, times=10)
```

```
## Call: survfit(formula = Surv(years, status2) ~ 1, data = dat.id)
##
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##    10     51     131    0.479  0.0359        0.413        0.554
```

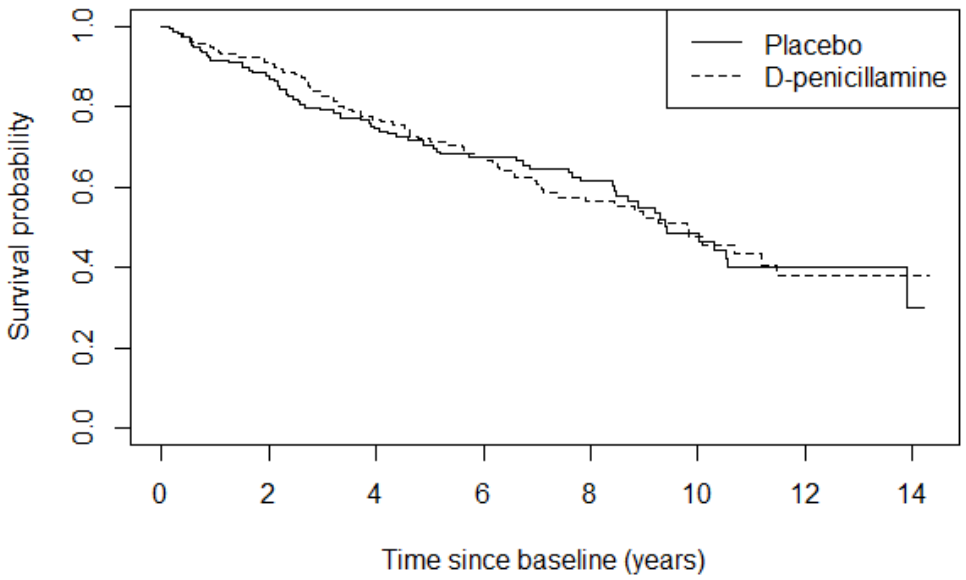
```
surv.pbc
```

```
## Call: survfit(formula = Surv(years, status2) ~ 1, data = dat.id)
##
##       n  events  median 0.95LCL 0.95UCL
##  312.00  140.00    9.43    8.68   11.17
```

Example: PBC data

- Compare survival curves across treatments

```
plot(survfit(Surv(years, status2) ~ drug, data=dat.id), ylab="Survival probability", xlab="Time s
ince baseline (years)",
     lty=c(1,2))
```



Example: PBC data

- Test for differences in survival curves between two groups: logrank test ("survdiff" function)
- Testing the null hypothesis of no difference in the hazard rates between the two groups
- P<0.05 indicates significant difference in the survival curves for the two groups

```
survdiff(Surv(years, status2) ~ drug, data=dat.id)
```

```
## Call:
## survdiff(formula = Surv(years, status2) ~ drug, data = dat.id)
##
##                 N Observed Expected (O-E)^2/E (O-E)^2/V
## drug=placebo   154       69     68.9  5.69e-05  0.000112
## drug=D-penicil 158       71     71.1  5.52e-05  0.000112
##
##  Chisq= 0  on 1 degrees of freedom, p= 1
```

Multivariable models

- Two broad categories

1. Proportional hazards approaches
   - Cox model (semi-parametric)
   - Fully parametric models
   - Assumes that the effect of the covariate is to multiply the hazard by some constant
   - E.g., for those in the treatment group the hazard of experiencing the event is 2 times that of those in the control group
2. Accelerated failure time models
   - Assumes that the effect of the covariate is to multiply the survival time by some constant
   - E.g., the time to the event is doubled for those in the treatment group versus those with the control group

Proportional hazards models

- The hazard is the instantaneous risk of the event given the subject is at risk
- Relative risk models assume a multiplicative effect of covariates on the hazard scale

$$h_i(t) = h_0(t)\exp(\gamma_1 w_{i1} + \gamma_2 w_{i2} + \ldots + \gamma_p w_{ip})$$

$$\log h_i(t) = \log h_0(t) + \gamma_1 w_{i1} + \gamma_2 w_{i2} + \ldots \gamma_p w_{ip}$$

- $h_i(t)$: hazard of an event for patient i at time t
- $h_0(t)$: baseline hazard (all covariates equal 0)
- $w_{i1}, \ldots, w_{ip}$: set of covariates
- $\gamma_1, \ldots, \gamma_p$: regression coefficients (log hazard ratios)

Proportional hazards models

- The baseline hazard $h_0(t)$ represents the instantaneous risk of experiencing the event at time t, without the influence of any covariate
- If a covariate has a beneficial effect, decreases $h_0(t)$: $\gamma < 0$
- If a covariate has a harmful effect, increases $h_0(t)$: $\gamma > 0$
- Hazard ratio (HR): $\frac{h_i(t|w_i)}{h_k(t|w_k)} = e^{\gamma'(w_i - w_k)}$
- HR > 1: event hazard increases (survival length decreases)
- HR < 1: event hazard decreases (survival length increases)
- Proportional hazards assumption

Proportional hazards model

- Corresponding cumulative hazard and survival functions

$$H(t|w) = H_0(t)e^{\gamma' w}, \qquad S(t|w) = S_0(t)^{\exp\{\gamma' w\}}$$

- where $H_0(t) = \int_0^t h_0(t)dt$ is the cumulative baseline hazard
- and $S_0(t) = e^{-H_0(t)}$ is the baseline survival function (survival function when all covariates are 0)
- HR=2: probability that a person in the treatment group will be alive at any given time t is the square of the probability that a person in the placebo group will be alive at the same time

Baseline hazard

- What is the form of the baseline hazard $h_0(t)$?
- Cox PH model: Does not specify a parametric form for the baseline hazard (semiparametric)
- Parametric PH model: Specifies a specific statistical distribution for the baseline hazard

Parametric PH models

- Example: Weibull distribution
- We begin with a Weibull baseline hazard: $h_0(t) = \lambda\alpha t^{\alpha-1}$
- Then the proportional hazards model is given by

$$h(t) = h_0(t)e^{\gamma' w} = (\lambda e^{\gamma' w})\alpha t^{\alpha-1}$$

- which is also a Weibull with $\lambda^* = \lambda e^{\gamma' w}$

$$h(t) = \lambda^*\alpha t^{\alpha-1}, \qquad S(t) = \exp(-\lambda^* t^\alpha)$$

- If $\alpha > 1$ then risk increases over time
- If $\alpha < 1$ then risk decreases over time

Parametric PH models

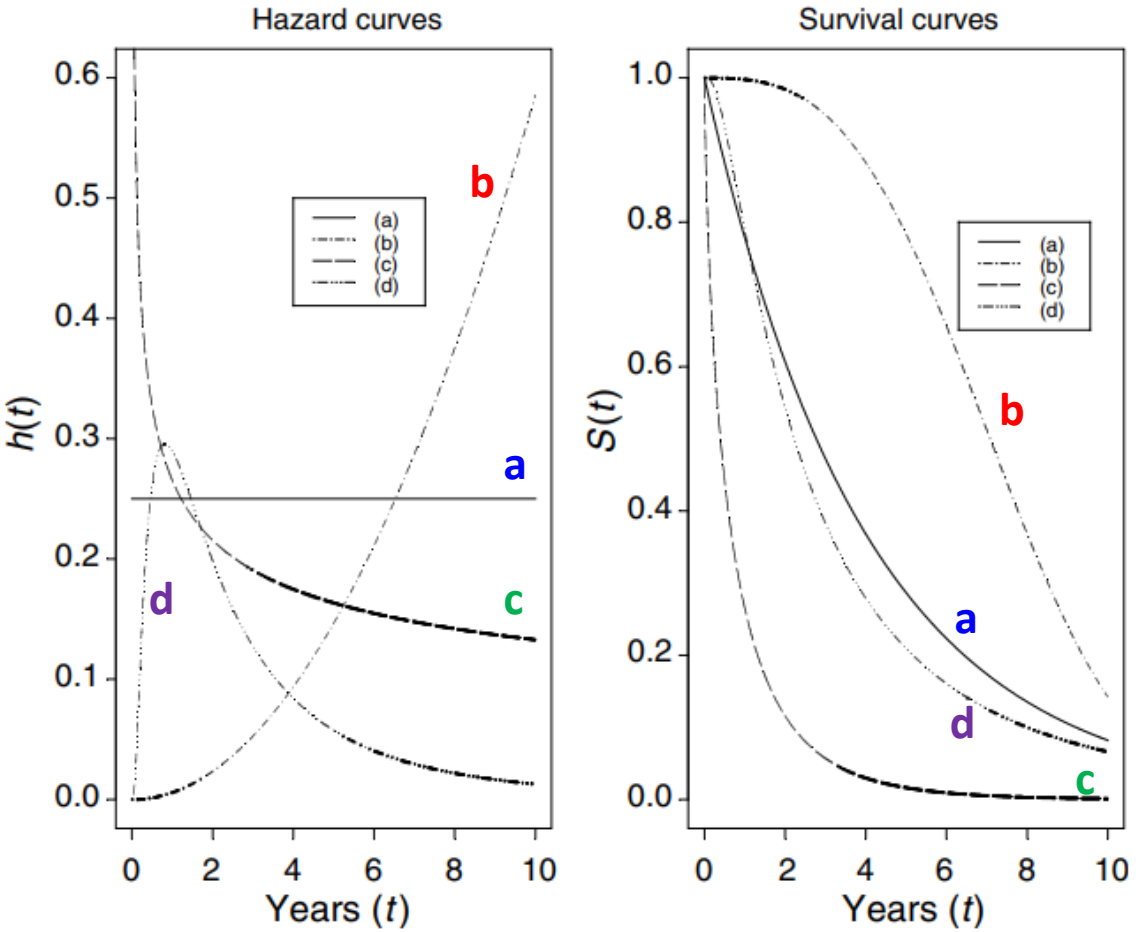


Figure 4 Relationships between (parametric) hazard and survival curves: (a) constant hazard (e.g. healthy persons), (b) increasing Weibull (e.g. leukaemia patients), (c) decreasing Weibull (e.g. patients recovering from surgery), (d) increasing and then decreasing log-normal (e.g. tuberculosis patients).

Clark et al. British Journal of Cancer (2003)

Parametric PH models

- Need to specify a distribution that matches the survival times (need to verify)
- Sometimes standard parametric distributions are not flexible enough to capture the shape of the underlying hazard function (e.g., Weibull is monotonic)
- Can make these models more flexible by considering piecewise splines for the time effect
- If a suitable distribution is found then can be easier to derive functions and obtain predictions
- Can also be slightly more efficient and yield more precise estimates

Likelihood for Censored Data

- Data: $\{T_i, \delta_i\}$
- $p(T_i; \theta)$ is the probability density function where

$$p(t) = h(t)S(t)$$

- Normally the log-likelihood would be given as

$$l(\theta) = \sum_{i=1}^{n} \log p(T_i; \theta)$$

- But we need to account for censoring

Likelihood for Censored Data

- Two types of contribution:

1. Subject who we observe an event for at time T_i ($\delta_i = 1$) contributes $p(T_i; \theta)$

2. Subject who is censored at time T_i ($\delta_i = 0$) we know that they have survived up to that point (i.e., $T_i^* > T_i = C_i$), so they contribute $S(T_i; \theta)$

$$l(\theta) = \sum_{i=1}^{n} \delta_i \log p(T_i; \theta) + (1 - \delta_i) \log S_i(T_i; \theta)$$

$$l(\theta) = \sum_{i=1}^{n} \delta_i \log h_i(T_i; \theta) - H_i(T_i; \theta)$$

Estimation of Relative Risk Models

- Sensitivity to distributional assumptions due to censoring
- Parametric proportional hazards model
  - Maximize the log-likelihood
- Cox (1972) showed that estimation can be conducted without specifying the baseline hazard
  - We make no assumptions for the baseline hazard function
  - Factor the baseline hazard out of the likelihood and maximize the resulting partial log-likelihood

Example: PBC data

- Fit a Cox PH model that contains main effects of drug, sex, and age

$$h(t) = h_0(t)\exp\{\gamma_1 \text{D-penicillin}_i + \gamma_2 \text{Female}_i + \gamma_3 \text{Age}_i\}$$

- Use "coxph" function in "survival" package

```
coxFit <- coxph(Surv(years, status2) ~ drug + sex + age, data=pbc2.id)
```

Example: PBC data

```
coxFit <- coxph(Surv(years, status2) ~ drug + sex + age, data=pbc2.id)
summary(coxFit)
```

```
## Call:
## coxph(formula = Surv(years, status2) ~ drug + sex + age, data = pbc2.id)
##
##   n= 312, number of events= 140
##
##                    coef exp(coef)  se(coef)      z Pr(>|z|)
## drugD-penicil -0.146013  0.864146  0.172143 -0.848   0.3963
## sexfemale     -0.470905  0.624437  0.221785 -2.123   0.0337 *
## age            0.042842  1.043773  0.008505  5.037 4.72e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##               exp(coef) exp(-coef) lower .95 upper .95
## drugD-penicil    0.8641     1.1572    0.6167    1.2109
## sexfemale        0.6244     1.6014    0.4043    0.9644
## age              1.0438     0.9581    1.0265    1.0613
##
## Concordance= 0.629  (se = 0.024 )
## Likelihood ratio test= 33.25  on 3 df,   p=3e-07
## Wald test            = 34.87  on 3 df,   p=1e-07
## Score (logrank) test = 35.31  on 3 df,   p=1e-07
```

Example: PBC data

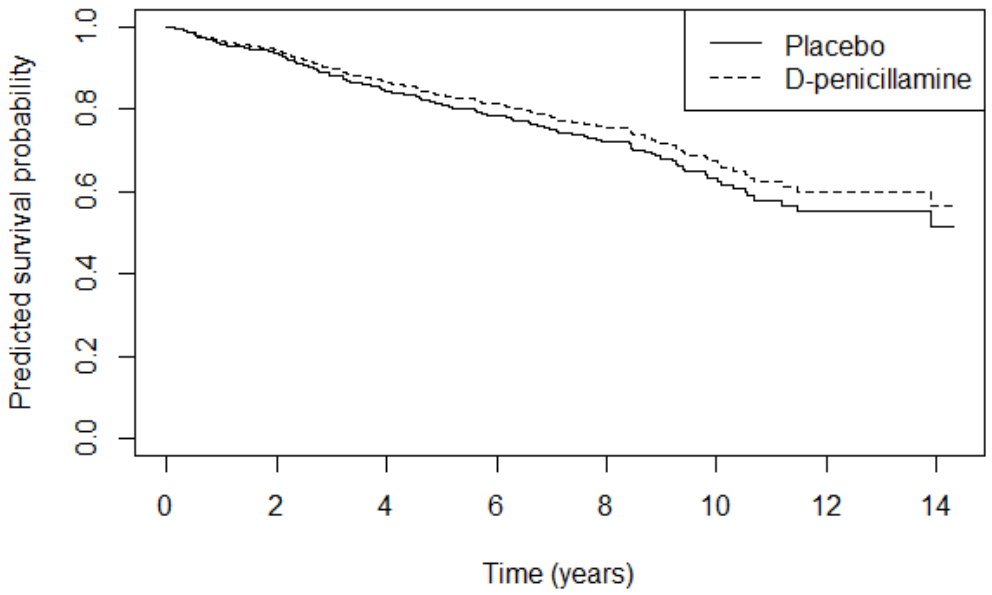
	exp(coef)	exp(-coef)	lower .95	upper .95
drugD-penicil	0.8641	1.1572	0.6167	1.2109
sexfemale	0.6244	1.6014	0.4043	0.9644
age	1.0438	0.9581	1.0265	1.0613

- There is no significant association between treatment and mortality risk (HR: 0.86; 95% CI: 0.62-1.21; p=0.40)
- The mortality risk for those in the treatment group is 14% lower than those in the placebo group
- This 14% relative decrease is independent of time
- Females have decreased risk of death (HR=0.62; 95% CI: 0.40-0.96; p=0.03)
- For every one year increase in age the mortality rate increases by 4% (HR=1.04; 95% CI: 1.03-1.06; p<0.001)

Example: PBC data

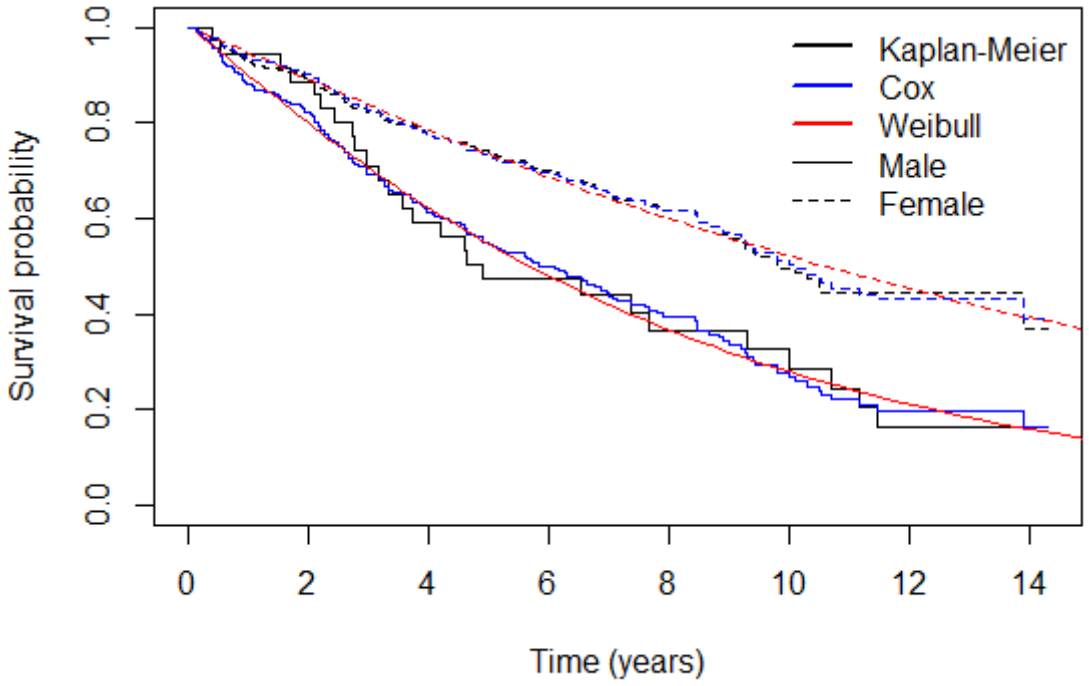
- Predict the survival curve from the Cox model if they received the drug or the placebo, for a new patient who is female and 40 years old at baseline

```
plot(survfit(coxFit, newdata=data.frame(drug="placebo",sex="female",age=40)), conf.int=FALSE)
lines(survfit(coxFit, newdata=data.frame(drug="D-penicil",sex="female",age=40)), conf.int=FALSE,
 lty=2)
legend("topright",c("Placebo","D-penicillamine"),lty=1:2)
```



Comparing models

- Can compare nested Cox models using likelihood ratio test
- Can compare Cox and parametric PH models using AIC



Breakout Session #3

1. What are two features of survival outcomes that require them to be analyzed with specific methods? What would be two consequences of applying standard regression methods?
2. When is the Kaplan-Meier estimate 0?
3. A clinical investigator wants to report the proportion of patients who experienced the event during the study. What is wrong with this estimate and what would you recommend reporting instead to describe the cohort's risk?
4. What are two benefits of using a parametric PH survival model compared to the Cox PH model?
5. What's your favourite animal?