JOHNS HOPKINS
BLOOMBERG SCHOOL
of PUBLIC HEALTH

Lecture 14

Continuous time survival analysis:
Review of survival function
Cox Proportional Hazards model

Review of Lecture 13

Let D be the time to an event of interest and Let C be the time to censoring, $D > 0$ and $C > 0$.

Define δ as the indicator that the event occurred ($\delta = 0$ if the event was censored).

Then, we get to observe $T_i = min(D_i, C_i)$ and δ_i for each subject i.

We assume D and C are independent and (T_i, δ_i) is independent of (T_j, δ_j) for all i and j.

Function	$F(t)$ Distribution	$S(t)$ Survival	$f(t)$ Density	$h(t)$ Hazard
Definition	$Pr(T \leq t)$	$Pr(T > t)$	$\lim\limits_{dt \to 0} \frac{Pr(t < T \leq t + dt)}{dt}$	$\lim\limits_{dt \to 0} \frac{Pr(t < T \leq t + dt \mid T > t)}{dt}$
Relationship to:				
$F(t)$		$1 - F(t)$	$\frac{d}{dt} F(T)$	$\frac{d}{dt} log\left(1 - F(T)\right)$
$h(t)$	$1 - exp\left(-\int_0^t h(u)du\right)$	$exp\left(-\int_0^t h(u)du\right)$	$h(t) exp\left(-\int_0^t h(u)du\right)$	

Targets of inference and why!

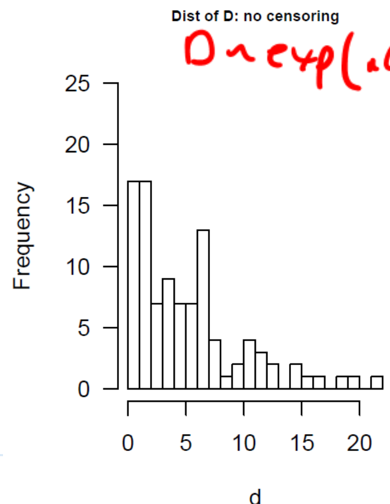
- We are primarily interested in making inference about
  - Survival function: S(t) = Pr(T > t)
  - Hazard function $$h(t) = \lim_{dt \to 0} \frac{Pr(t < T \leq t + dt | T > t)}{dt} = \frac{f(t)}{S(t)} = -\frac{d}{dt} log S(t)$$

- But why?
  - Censoring complicates the estimation procedures
  - Consider f(t), density with and without censoring

- Estimation / inference for S(t) and h(t) can "easily" incorporate censoring

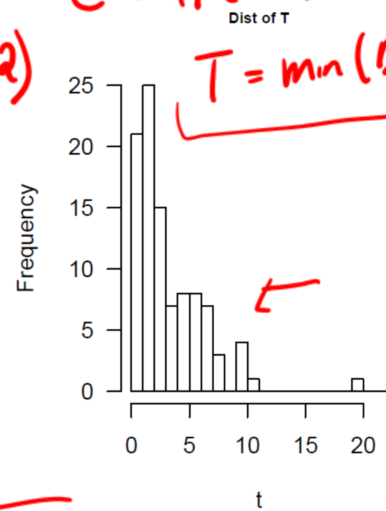
Pr(T > t)

D ~ exp(.02)



C ~ exp(.01)

T = min(D,C)



Kaplan-Meier estimate of the survival function

The Kaplan-Meier estimate of the survival function $S(t)$ is also known as the **Product-limit** estimator.

This estimator for the survival function assumes that:

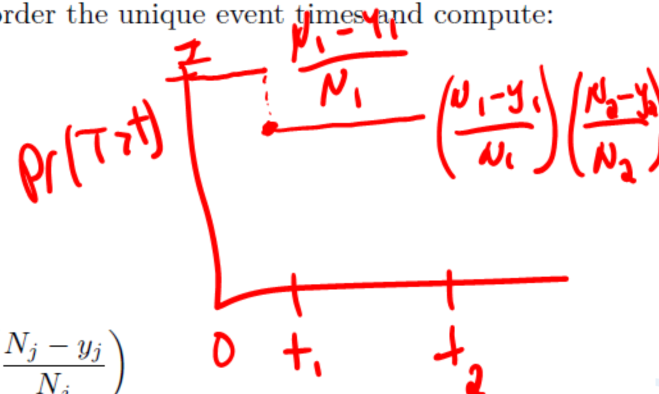
- censoring is unrelated to prognosis, i.e. event process and censoring process are independent
- the survival probabilities are the same for subjects recruited early and late in the study
- the events happened at the times specified

To construct the Kaplan-Meier estimator, you need to order the unique event times and compute:

Event times:	t_1	$<$	t_2	$<$	...	$<$	t_J
No. at risk:	N_1	$>$	N_2	$>$	...	$>$	N_J
No. of events:	y_1		y_2		...		y_J

The estimate of $S(t)$ is 1 if $t < t_1$ and

$$\hat{S}(t) = \prod_{j:t_j \leq t} \left(\frac{N_j - y_j}{N_j} \right)$$



Kaplan-Meier estimate of the survival function

- Using the data for inpatients hospitalized for a severe mental disorder, we will be computing the Kaplan-Meier estimate of the survival function for the female patients. Survival time from hospitalization is in years.

year of hospitalization

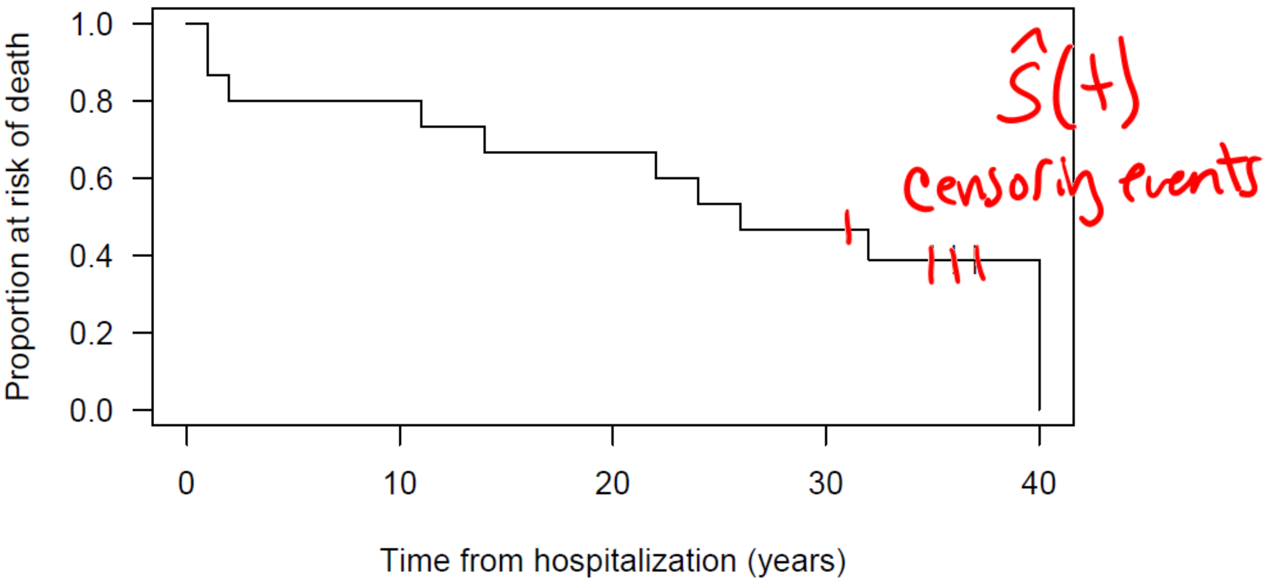
observed follow-up

at risk

of events

	1	1	2	11	14	22	24	26	31+	32	35+	35+	36+	37+	40	Ni	yi	(Ni-yi)/Ni	S(t)
1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	15	2	0.867	0.867
2			1	0	0	0	0	0	0	0	0	0	0	0	0	13	1	0.923	0.800
3				0	0	0	0	0	0	0	0	0	0	0	0	12	0	1.000	0.800
4				0	0	0	0	0	0	0	0	0	0	0	0	12	0	1.000	0.800
5				0	0	0	0	0	0	0	0	0	0	0	0	12	0	1.000	0.800
6				0	0	0	0	0	0	0	0	0	0	0	0	12	0	1.000	0.800
7				0	0	0	0	0	0	0	0	0	0	0	0	12	0	1.000	0.800
8				0	0	0	0	0	0	0	0	0	0	0	0	12	0	1.000	0.800
9				0	0	0	0	0	0	0	0	0	0	0	0	12	0	1.000	0.800
10				0	0	0	0	0	0	0	0	0	0	0	0	12	0	1.000	0.800
11				1	0	0	0	0	0	0	0	0	0	0	0	12	1	0.917	0.733
12					0	0	0	0	0	0	0	0	0	0	0	11	0	1.000	0.733
13					0	0	0	0	0	0	0	0	0	0	0	11	0	1.000	0.733
14					1	0	0	0	0	0	0	0	0	0	0	11	1	0.909	0.667

Kaplan-Meier estimate of survival function



Greenwood's formula for variance of S(t)

An estimate of the variance of $\hat{S}(t)$ based on Greenwood's formula (application of Delta method) is:

$$\hat{V}ar(\hat{S}(t)) = \hat{S}(t)^2 \sum_{j:t_j \leq t} \frac{y_j}{N_j(N_j - y_j)}$$

A 95% confidence interval for $S(t)$ can be derived as:

$$\hat{S}(t) \pm 1.96\sqrt{\hat{V}ar(\hat{S}(t))}$$

with imposing the constraint that the confidence interval lies in $[0, 1]$, i.e. if the bounds of the confidence interval go outside $[0, 1]$, set the values to 0 or 1, respectively. This is unappealing in many respects!

Variance of S(t) estimate based on complementary log-log

An alternative to Greenwoods formula for the variance, a variance estimate can be derived based on the complementary Log-Log transformation.

Let $v(t) = log[-logS(t)]$. Note that $S(t) \in [0,1]$ and $v(t) \in [-\infty, \infty]$.

$$\hat{V}ar(\hat{v}(t)) = \sum_{j:t_j \leq t} \frac{y_j}{N_j(N_j - y_j)} \left[\sum_{j:t_j \leq t} log\left(\frac{N_j - y_j}{N_j}\right) \right]^{-2}$$

The 95% confidence interval for $v(t)$ is given by:

$$\hat{v}(t) \pm 1.96\sqrt{\hat{V}ar(\hat{v}(t))}$$

where we can define the upper and lower bound as $\hat{v}_L(t)$ and $\hat{v}_U(t)$.

NOTE: $S(t) = exp(-exp(v(t)))$, so the 95% confidence interval for $S(t)$ is:

$$[exp(-exp(\hat{v}_U(t))), exp(-exp(\hat{v}_L(t)))]$$

Example calculations: Greenwood's formula

Compute the 95% confidence interval for $S(2)$:

1. Using Greenwood's formula:

$$
\begin{aligned}
\hat{V}ar(\hat{S}(2)) &= \hat{S}(2)^2 \sum_{j:t_j \leq 2} \frac{y_j}{N_j(N_j - y_j)} \\
&= \hat{S}(2)^2 \left[\frac{y_1}{N_1(N_1 - y_1)} + \frac{y_2}{N_2(N_2 - y_2)} \right] \\
&= 0.8^2 \left[\frac{2}{15 \times (15-2)} + \frac{1}{13 \times (13-1)} \right] \\
&= 0.0107
\end{aligned}
$$

95% CI for $S(2)$: $0.8 \pm 1.96 * \sqrt{0.0107} \rightarrow (0.598, 1.003)$

Example calculations: Complementary log-log

2. Using the Complementary Log-Log transformation

$$
\begin{aligned}
\hat{v}(2) &= log(-log(\hat{S}(2))) \\
&= log(-log(0.8)) \\
&= -1.50 \\
\hat{V}ar(\hat{v}(2)) &= \sum_{j:t_j\le 2} \frac{y_j}{N_j(N_j-y_j)} \left[\sum_{j:t_j\le 2} log\left(\frac{N_j-y_j}{N_j}\right)\right]^{-2} \\
&= \left[\tfrac{y_1}{N_1(N_1-y_1)} + \tfrac{y_2}{N_2(N_2-y_2)}\right]\left[log\left(\tfrac{N_1-y_1}{N_1}\right) + log\left(\tfrac{N_2-y_2}{N_2}\right)\right]^{-2} \\
&= \left[\tfrac{2}{15\times 13} + \tfrac{1}{13\times 12}\right][log(13/15) + log(12/13)]^{-2} \\
&= 0.335
\end{aligned}
$$

95% CI for $v(2)$ is: $\hat{v}(2) \pm 1.96\sqrt{\hat{V}ar(\hat{v}(2))}$ is $-1.50 \pm 1.96\sqrt{0.335}$ is $(-2.63, -0.36)$.

95% CI for $S(2)$ is: $(exp(-exp(-0.36)), exp(-exp(-2.63)))$ is $(0.50, 0.93)$.

R implementation

```
library(survival)
St.green = survfit(Surv(survive,event) ~ 1, data = d.female,
                   type = "kaplan-meier",
                   conf.type = "plain")
St.cll = survfit(Surv(survive,event) ~ 1, data = d.female,
                   type = "kaplan-meier",
                   conf.type = "log-log")
summary(St.green)
```

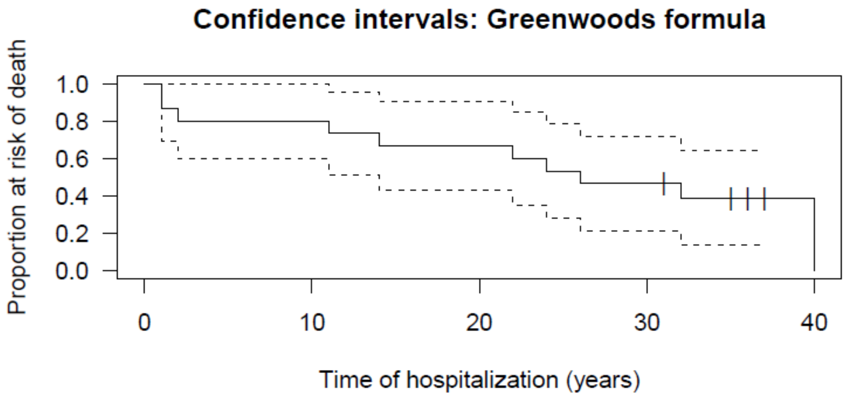
T_i δ_i

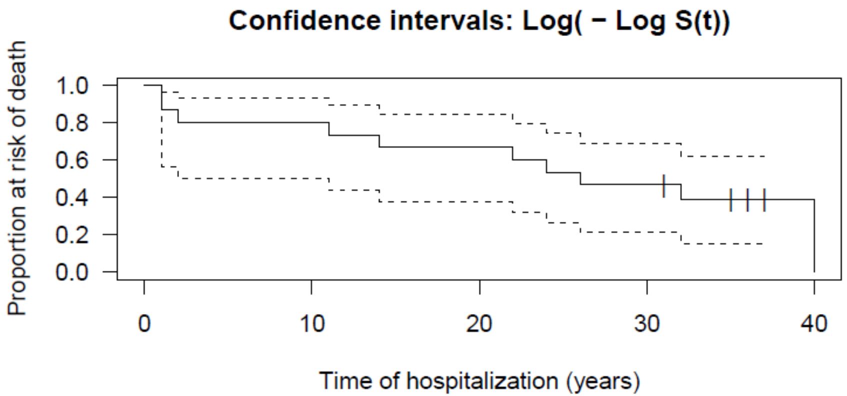
greenwoods formula

```
## Call: survfit(formula = Surv(survive, event) ~ 1, data = d.female,
##     type = "kaplan-meier", conf.type = "plain")
##
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##     1     15       2    0.867  0.0878        0.695        1.000
##     2     13       1    0.800  0.1033        0.598        1.000
##    11     12       1    0.733  0.1142        0.510        0.957
##    14     11       1    0.667  0.1217        0.428        0.905
##    22     10       1    0.600  0.1265        0.352        0.848
##    24      9       1    0.533  0.1288        0.281        0.786
##    26      8       1    0.467  0.1288        0.214        0.719
##    32      6       1    0.389  0.1287        0.137        0.641
##    40      1       1    0.000     NaN          NaN          NaN
```

$\| \hat{S}(t)$

R implementation





Log rank test

Regression models for the hazard function

- Most famous and commonly used model: Cox proportional hazards model

building a model for log hazard

semi-parametric regression model

$$\lambda(t|X) = \lambda_0(t)e^{X\beta}$$

$h(t) = \lambda(t)$

$$\Rightarrow \quad log(\lambda(t|X)) = log(\lambda_0(t)) + X\beta$$

willing to assume a distribution for $D \sim$ exp gamma

$\downarrow$

parametric regression models

where

- $X = (X_1, X_2, ..., X_p)$, no intercept!
- $log(\lambda_0(t))$ is the "baseline hazard" and is the intercept which depends on t
- $\beta_j = log\left(\frac{\lambda(t|X_1,...,X_j=x_j+1,...,X_p)}{\lambda(t|X_1,...,X_j=x_j,...,X_p)}\right)$, the log relative hazard.

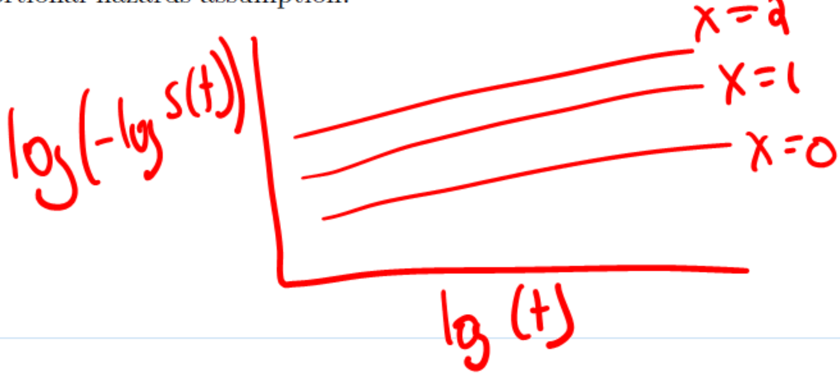
Exploring the proportional hazards assumption

Recall that:

$$
\begin{aligned}
S(t|X) &= exp\left(-\int_0^t \lambda_0(u)e^{X\beta}du\right) \\
&= exp\left(-e^{X\beta}H_0(t)\right),\ H_0(t) \text{ is the baseline cumulative hazard}
\end{aligned}
$$

$$\boxed{log(-log(S(t|X)))} = log(H_0(t)) + X\beta$$

So plotting the $log(-log(S(t|X = x)))$ vs. $log(t)$ for values of x will allow you to visually inspect the proportional hazards assumption. If $log(-log(S(t|X = x)))$ for different values of x are parallel, then this supports the proportional hazards assumption.



Parallel
⇓
proportional hazards hold

Estimation within Cox model

$\Rightarrow \beta s$

- Estimation for association parameters for an arbitrary baseline hazard
  - Utilizes a partial or profile likelihood approach

$\to \log \lambda_0(t)$ nuisance parameter

The estimation procedures maximizes the partial likelihood function for event times $t_1 < t_2 < ... < t_n$ with risk sets (i.e. subjects who are still at risk of experiencing the events) $R_1 \supset R_2 \supset .... \supset R_n$.

$$
\begin{aligned}
L(\beta) &= \prod_{i=1}^{n} Pr(\text{person } i \text{ has the event at } t_i | 1 \text{ person in risk set } R_i \text{ has the event}) \\
&= \prod_{i=1}^{n} \left[\frac{\boxed{\lambda_0(t) e^{X_i'\beta}}}{\sum_{j \in R_i} \lambda_0(t) e^{X_j'\beta}} \right] \\
&= \prod_{i=1}^{n} \left[\frac{e^{X_i'\beta}}{\sum_{j \in R_i} e^{X_j'\beta}} \right]
\end{aligned}
$$

Profile likelihood

β_s $\lambda_0(t)$

Assume you have the following model: $Pr(Y = y|\theta = (\theta_1, \theta_2))$ where θ_1 is of interest, and θ_2 is a nuisance (it is an unknown but you are not interested in making inference about it).

- You observe $y_1, ..., y_n$ and the likelihood function is: $L(y|\theta) = \prod_{i=1}^{n} f(y_i|\theta)$.
- Define $\hat{\theta}_2(\theta_1, y)$ to be the value for $\hat{\theta}_2$ that maximizes the likelihood (solves the score equation) when θ_1 is fixed.
- The profile likelihood is then defined as $PL(y|\theta) = \prod_{i=1}^{n} f(y_i|\theta_1, \hat{\theta}_2)$.
- If $\hat{\theta}_1$ maximizes the profile likelihood, then it is the maximum likelihood estimate.

- The idea is that we estimate the baseline hazard using a non-parametric estimate, then estimate the association parameters assuming the baseline hazard is known/fixed.

Example

Going back to the example of time to death from hospitalization among a group of persons hospitalized for a severe mental disorder.

We will consider two Cox Proportional Hazards models:

- $\boxed{\text{Model A: } log(\lambda(t|male)) = log(\lambda_0(t)) + \beta_1 male}$
- Model B: $log(\lambda(t|male, age)) = log(\lambda_0(t)) + \beta_1 male + \beta_2 age$

binned survival data: model C and D assumed PH

```
library(survival)
d = read.table("./survival.csv",sep=",",header=T)
d$event = 1 - d$censor

fitA = coxph(Surv(survive,event)~male,data=d)
summary(fitA)
```

T_i, δ_i

```
## Call:
## coxph(formula = Surv(survive, event) ~ male, data = d)
##
##   n= 26, number of events= 14
##
##         coef exp(coef) se(coef)      z Pr(>|z|)
## male -0.7511    0.4718   0.6055 -1.241    0.215
##
##      exp(coef) exp(-coef) lower .95 upper .95
## male    0.4718      2.119     0.144     1.546
```

β_{male}

At any year of hospitalization the hazard of death for males is 53% smaller than the hazard of death for females

Example

- Model B: $log(\lambda(t|male, age)) = log(\lambda_0(t)) + \beta_1 male + \beta_2 age$

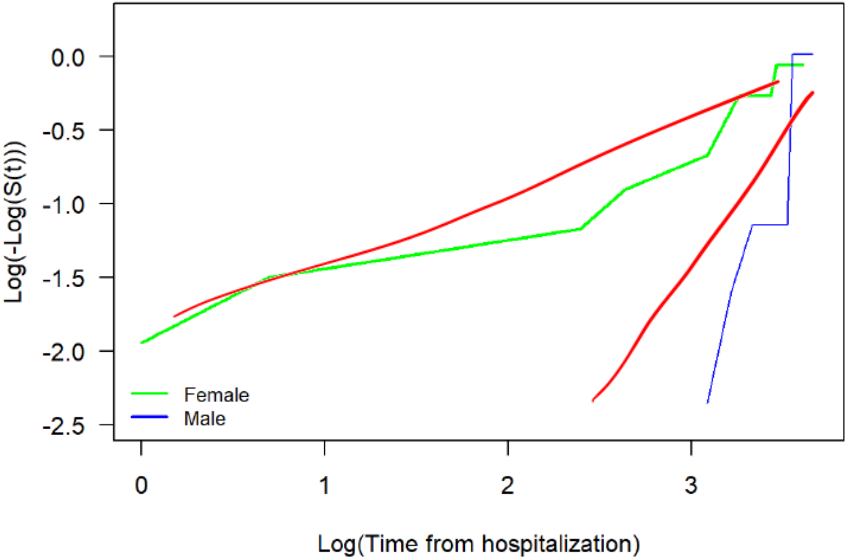
```r
fitB = coxph(Surv(survive,event)~male+age,data=d)
summary(fitB)
```

```
## Call:
## coxph(formula = Surv(survive, event) ~ male + age, data = d)
##
##   n= 26, number of events= 14
##
##         coef exp(coef) se(coef)     z Pr(>|z|)
## male 0.52374   1.68833  0.73753 0.710  0.47762
## age  0.20753   1.23063  0.05828 3.561  0.00037 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##      exp(coef) exp(-coef) lower .95 upper .95
## male     1.688     0.5923    0.3978     7.165
## age      1.231     0.8126    1.0978     1.380
```

At any year of hospitalization, the hazard of death for males is 69% greater than that for females, among persons hospitalized at the same age.

Checking the proportaional hazards assumption

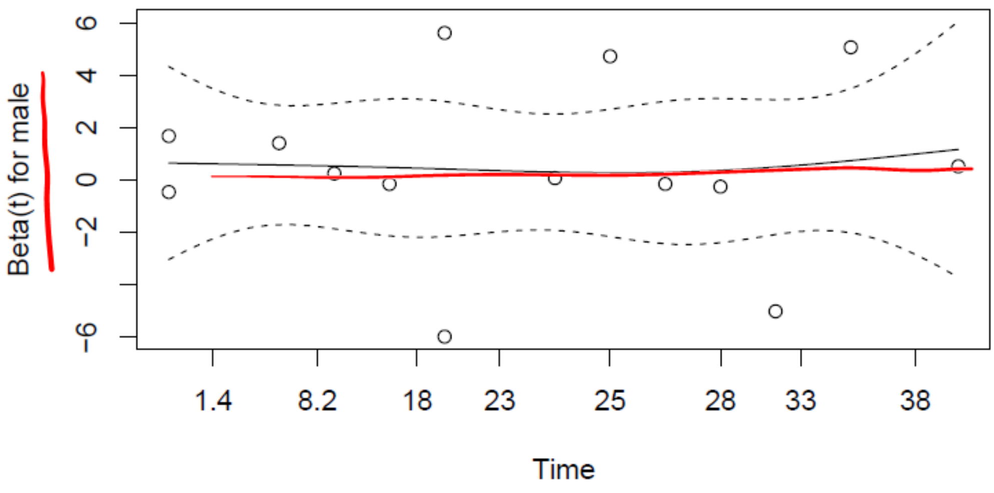
- Plot the log(- log(S(t)) as a function of log(t) separately for males and females



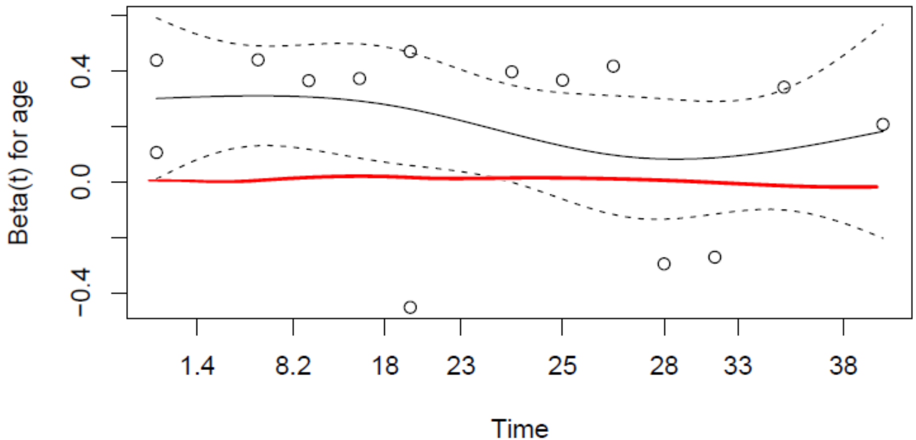
Alternative evaluation of the proportional hazards assumption

- Schoenfeld residuals plot.
  - If mean residuals are 0 across time, then proportional hazards assumption holds
  - The x-axis takes the unique event times and plots these scaled to the estimates of S(t)

```
temp <- cox.zph(fitB)
par(mfrow=c(2,1),mar=c(4,4,1,1))
plot(temp)
```



Alternative evaluation of the proportional hazards assumption



- Looks like there is a violation of the proportional hazards assumption for age
  - There are ways to account for non-proportional hazards, e.g. estimate a time specific effect of age.

- Here is one vignette that is a good starting place:

https://cran.r-project.org/web/packages/Greg/vignettes/timeSplitter.html

Next time

- Special topic:
  - Post a lecture on adjustment for multiple comparisons (last year)
  - Variable selection procedures for baseline covariate adjustment to improve precision of marginal treatment effects in randomized trials

RF, lasso

prediction
Cox model
machine learning
RF

→ regression/risk factor analysis
→ S(t) estimates
→ Cox model