Survival analysis: Coursework

Petteri Mäntymaa

May 20, 2019

Q1

a)

Show that ***the hazard function*** *is uniquely determined by the* ***mean residual time*** $r(t)$.

Mean residual time $r(t)$ at time t can be defined as

$$r(t) = E(X - t \mid X \geq t) = \frac{\int_t^\infty (u-t)f(u)du}{S(t)}$$

where $f(t)$ is the density function.

We know that the continuous survival function with a finite mean is uniquely determined by the mean residual time as

$$S(t) = \frac{r(0)}{r(t)} \exp\Big(- \int_0^t \frac{1}{r(u)} du \Big)$$

Thus, as the hazard function $\lambda(t)$ can be determined by the survival function

$$\lambda(t) = \frac{f(t)}{S(t)} = \frac{-\frac{d}{dt}S(t)}{S(t)} = -\frac{d}{dt}\log(S(t))$$

Where $\frac{d}{dt}S(t) = \frac{d}{dt}1 - F(t) = -\frac{d}{dt}F(t) = -f(t)$.

$\lambda(t)$ is uniquely determined by $r(t)$ by substituting $S(t)$ as

$$\begin{aligned}
-\frac{d}{dt}\log(S(t)) &= -\frac{d}{dt}\log\left(\frac{r(0)}{r(t)} \exp\Big(- \int_0^t \frac{1}{r(u)} du \Big)\right) \\
&= -\frac{d}{dt}\Big(\log(r(0)) - \log(r(t)) - \int_0^t \frac{1}{r(u)} du \Big) \\
&= -\frac{d}{dt}\Big(\log(r(0)) \Big) + \frac{d}{dt}\Big(\log(r(t)) \Big) + \frac{d}{dt}\Big(\int_0^t \frac{1}{r(u)} du \Big) \\
&= -(0) + \frac{\frac{d}{dt}r(t)}{r(t)} + \frac{1}{r(t)}dt \\
&= \frac{\frac{d}{dt}r(t) + 1}{r(t)}
\end{aligned}$$

b)

Further, show that the exponential distribution is the only continuous distribution for which the mean residual lifetime $r(t)$ *is constant for all* $t > 0$.

We know that the mean residual time $r(t)$ can also be expressed as

$$r(t) = E(X - t \mid X \geq t) = \frac{\int_t^\infty S(u)du}{S(t)}$$

Thus, with the exponential survival function for $t \geq 0$ we get

$$\begin{aligned} r(t) = \frac{\int_t^\infty S(u)du}{S(t)} &= \frac{\int_t^\infty \exp(-\lambda u)du}{\exp(-\lambda t)} \\ &= \frac{\frac{1}{\lambda}\exp(-\lambda t)}{\exp(-\lambda t)} \\ &= \frac{1}{\lambda} \end{aligned}$$

Which of course is the mean of the exponential distribution $E(T)$. This means that the residual life $(X - t) \mid X > t$ has the same distribution as X. This is because the memorylessness property of the exponential distribution which states that for all $x, t \geq 0$

$$P(X > t + x \mid X > x) = P(X > t)$$

In our case

$$\begin{aligned} P(X > t + x \mid X > x) &= \frac{S(t+x)}{S(x)} \\ &= \frac{\exp(-\lambda(t+x))}{\exp(-\lambda x)} \\ &= \exp(-\lambda t) = P(X > t) \end{aligned}$$

Q2

a)

Define at-risk process and counting process for time-to-failure data.

Let $X_1, \ldots, X_n$ be a sample of uncensored continuously distributed survival times. Denote survival function at time t as $S(t) = P(X > t)$ and hazard as as $\alpha(t)dt = P(t \leq X < t + dt \mid X \geq t)$. Thus, naturally the cumulative hazard is $A(t) = \int_0^t \alpha(s)ds$. As we know the hazard function defines the survival function by $\exp(-\int_0^t \alpha(s)ds)$.

The data itself can be defined as

$$(\tilde{X}_i, D_i) = \begin{cases} X_i = \tilde{X}_i, & D_i = 1 \\ X_i > \tilde{X}_i, & D_i = 0 \end{cases}$$

for $i = 1, \ldots, n$, where D_i denotes censoring for individual i. When $D_i = 1$, $\tilde{X}_i$ is the survival time and when $D_i = 0$, $\tilde{X}_i$ is the censoring time.

Let $\mathcal{F}_t$ denote filtration, as in everything that is known at time t. This means that $\mathcal{F}_t$ is a set of realisations of random variables prior to and including t. $\mathcal{F}_t-$ denotes what is known just prior to but not including t. I guess $\mathcal{F}_t-$ could be thought of as a limit $\lim\limits_{\epsilon \to 0} \mathcal{F}_{t-\epsilon}$.

We can condition on filtration to derive, what resembles a delta function. For the hazard rate we get

$$P(t \leq \tilde{X}_i < t + dt,\, D_i = 1 \mid \mathcal{F}_t-) = \begin{cases} \alpha(t)dt, & \tilde{X}_i \geq t \\ 0, & \tilde{X}_i < t \end{cases}$$

So the expectation for the sum over n indicator functions $1\{t \leq \tilde{X}_i < t + dt,\, D_i = 1\}$ given $\mathcal{F}_t-$ yields what can be described the *intensity.*

$$\begin{aligned} E(\sum_{i=1}^{n} 1\{t \leq \tilde{X}_i < t + dt,\, D_i = 1\} \mid \mathcal{F}_t-) &= \sum_{i=1}^{n} 1\{\tilde{X}_i \geq t\}\alpha(t)dt \\ &= \sum_{i=1}^{n} Y_i(t)\,\alpha(t)dt \\ &= Y(t)\,\alpha(t)dt \\ &= \lambda(t)dt \end{aligned}$$

where $Y_i(t)$ for individual i is 1 if i is at risk at time t and is 0 otherwise, thus the intensity for individual i is the hazard rate when the individual is at risk and zero when the event has happened. $Y(t)$ then denotes the size of the risk set at time t.

This defines the *at-risk process for time-to-failure data.*

We then define the *counting process for time-to-failure data*, that is, has an event happened prior to or at t.

As $Y_i(t) = 1\{\tilde{X}_i \geq t\}$ denotes the individuals status on being at risk at time t, a counting process

$$N_i(t) = 1\{\tilde{X}_i \leq t, \quad D_i = 1\}, \quad 0 \leq t \leq \tau$$

denotes whether the event has happened prior to or at t or not.

An increment of $N_i(t)$ is $dN_i(t) = N_i(t) - N_i(t-)$ and, as $N_i(t)$ is an indicator function, maps to either 1 on event or 0 otherwise. For a cohort of n individuals $\sum_{i=1}^{n} N_i(t) = N(t)$ denotes the failure counting process.

b)

What is the expectation of the increment conditional on the history of the process.

An increment over a small interval $[t, t + dt)$ is

$$dN(t) = N((t + dt)-) - N(t-) = \sum_{i=1}^{n} 1\{t \leq \tilde{X}_i < t + dt, \quad D_i = 1\}$$

thus the expected value of increment given the history (filtration prior to t) is, as discussed earlier in this exercise,

$$\begin{aligned} E(dN(t) \mid \mathcal{F}_t-) &= E(\sum_{i=1}^{n} 1\{t \leq \tilde{X}_i < t + dt, \quad D_i = 1\} \mid \mathcal{F}_t-) \\ &= \sum_{i=1}^{n} 1\{\tilde{X}_i \geq t\}\alpha(t)dt \\ &= \sum_{i=1}^{n} Y_i(t)\,\alpha(t)dt \\ &= Y(t)\,\alpha(t)dt \\ &= \lambda(t)dt \end{aligned}$$

Data

Consider a study of 30 cases with complex pulmonary atresia collected as the basis of an observational study on the presentation and natural history of this disease.

The variables are descrined below.

- agepres: age at presentation
- agelast: age last seen alive (if alive) or age at death (if dead)
- ageopl: age at first operation (-1 for no operation to date)
- dead: no=0, yes=1
- sex: male=0, female=1
- paanat: size of intrapericardial pulmonary arteries at presentation: absent or tiny=0, normal or near normal=1
- adfol: adequate follow-up. Study closed less than 1 year since presentation=0, study closed at least 1 year since presentation=1
- Follow.up: duration of follow-up (agelast-agepres)

Q3

Read the data as a data frame:

```
cpa <- read.csv(file = "complex-pulmonary-atresia-data.csv", sep = ";")
```

a)

How many patients completed the follow-up of one year and how many did not?

The variable `Follow.up` represents the follow up time in days from age at presentation to age at death (event) or last seen alive (censoring).

Let's produce a table how many of the patients had a follow time of at least a year:

```
table(cpa$Follow.up >= 365)
```

```
##
## FALSE  TRUE
##    13    17
```

17 patients completed the one-year follow up and 13 patients did not.

b)

What is the size of the risk set at 1 year?

The risk set is the number of patients still being followed at 1 year, that is not yet experienced an event or censored.

```
sum(cpa$Follow.up >= 365)
```

```
## [1] 17
```

The size of the risk set is 17.

Q4

To answer the question related to the survival experience of all patients from presentation, specify the following.

- inclusion criteria
  - Inclusion criteria are the properties and characteristics that an individual must have to be included in a study.
- outcome
  - Outcome is the individuals **time to event**, as in time from entry to the occurrence of an event.
- time origin
  - Time origin of a study is the time regarded as $T = 0$. For example birth, time of diagnosis or the time of entry to the study. Basically any fixed date or time.
- entry time
  - Entry time, as hinted above, is the time of entry to the study. Individuals aren't always followed from time 0. It can then be regarded as delayed entry or left truncation.
- censoring rule
  - answer
- survival time
  - Survival time is the time an indiviudal spends at risk before an event.
- period of observation
  - Period of observation is the total time span that observations are gathered in a study.

Q5

Estimate the survival function from presentation and plot the estimate with its confidence interval.

We'll calculate a Kaplan-Meier estimate of the survival function. It is a non-parametric method, thus no assumptions are made about how the hazard rate varies in time. The method treats time as a succession of small *clicks*, each of which (should) contain only one event at most. Let n be the initial risk set, d_j the number of events at time t_j and l_j the number of censorings at time t_j.

Thus the risk set at time t_i is given by

$$y_i = n - \sum_{j=0}^{i-1} d_j - \sum_{j=0}^{i-1} l_j$$

The conditional survival probability given d_i events at time (click) t_i is $p(T > t_i \mid T > t_{i-1}) = 1 - \frac{d_i}{y_i}$ so the Kaplan-Meier estimator estimates the probability to survive beyond time t by estimating surviving a succession of the previous event times (clicks).

$$\hat{S}(t) = P(T > t) = \prod_{i:t_i \leq t} \left(1 - \frac{d_i}{y_i}\right)$$

We'll estimate the survival function using the `survfit` function from time of presentation `agepres` to time last seen alive `agelast`. `dead` yields the status for known events. We will also draw 95 % confidence intervals which, with `conf.type="plain"`, are given by

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

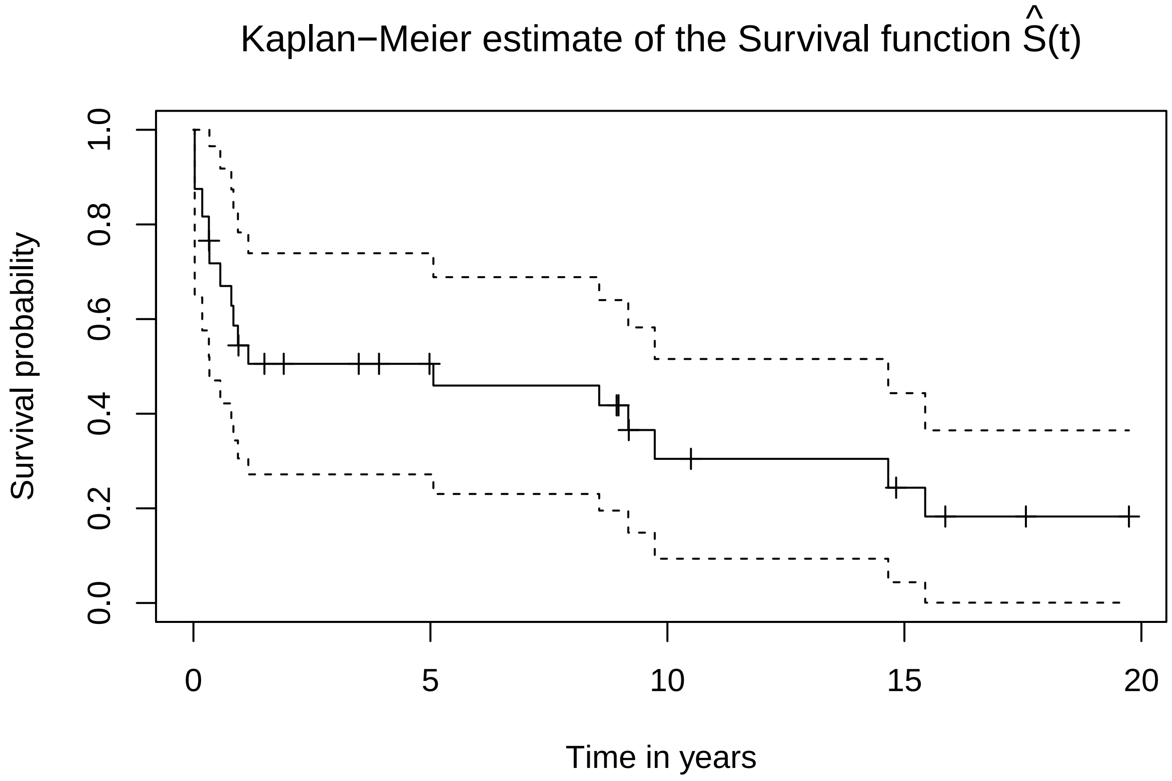
where $Var(\hat{S}(t)) = (\hat{S}(t))^2 \sum_{j:t_j \leq t} \frac{d_j}{y_j - d_j}$ is the so called Greenwood formula. Other options are the `conf.type="log"` where the survival function is scaled by taking the log, as $\log(\hat{S}(t))$ or the complimentary log-log, as $\log(-\log(\hat{S}(t)))$ with `conf.type="log-log"`.

As a thumb rule these are used if the confidence upper limit exceeds 1 (log-log) or the lower is less than 0 (log), but I fail to understand the precise intepretation and reprecautions of the log and log-log transformations in regards to the intepretation of the confidence limits. Hence, we use `plain`.

```r
cpa.km <- survfit(Surv(time = agepres, time2 = agelast, event = dead) ~ 1,
                  conf.int = 0.95,
                  conf.type = "plain",
                  type = "kaplan-meier",
                  data = cpa)
```

Below is the survival plot of the Kaplan-Meier estimate. The time is scaled to years for convenience. The tick-marks represent censoring.

```r
plot(x = cpa.km, mark.time = T, xscale = 365.25, xlab = "Time in years", ylab = "Survival probability",
```



The survival probability plummets in the first year and then continues to decline but significantly less rapidly. Censoring seems quite random at least by visual inspection. Note that the survival curve does not reach zero, more about that below.

a)

What are the sizes of risk sets in the beginning and at the end?

The follow up time `Follow.up` in ascending order standardizes the time scale to start from 0 for all patients (a row of time 0 must be added). The follow up is known for all patients, thus the follow up time is the time to event if event indicator `dead` is 1 and the time to censoring otherwise. The number of patients included in the study is as well known.

We know the risk set at time t to be

$$y_i = n - \sum_{j=0}^{i-1} d_j - \sum_{j=0}^{i-1} l_j$$

As there is one row of data per patient, the number of rows is the initial risk set size n. This can be verified via unique patient id's

```r
identical(n_distinct(unique(cpa$Patient)),
          (nrow(cpa)),
          (length(cpa$Patient)))
```

```
## [1] TRUE
```

In addition we calculate the cumulative number of events and the cumulative number of censorings at time t. The cumulative censorings are given by the complement of having experienced an event at time t. Finally add the "zero" row.

```r
cpa <- cpa %>%
  arrange(Follow.up) %>%
  mutate(n = n(),
         cum_dead = cumsum(dead),
         cum_censored = cumsum(1-dead),
         cum_y = n - cum_dead - cum_censored) %>%
  add_row(Follow.up = 0, n = 30, cum_dead = 0, cum_censored = 0, cum_y = n, .before = 1)
```

This provides what is basically a life table, but with cumulative values.

The first 6 rows:

```r
cpa %>%
  select(Patient, Follow.up, cum_y, cum_dead, cum_censored) %>%
  head()
```

```
##   Patient Follow.up cum_y cum_dead cum_censored
## 1      NA         0    30        0            0
## 2       5         4    29        1            0
## 3      15        14    28        2            0
## 4      21        77    27        3            0
## 5      19        88    26        4            0
## 6      22        89    25        4            1
```

The size of the risk set `cum_y` is 30 at the beginning.

```r
cpa %>%
  select(Patient, Follow.up, cum_y, cum_dead, cum_censored) %>%
  tail()
```

```
##    Patient Follow.up cum_y cum_dead cum_censored
## 26      18      3167     5       15           10
## 27        7       3254      4        15             11
## 28        6       5409      3        15             12
## 29       13       5759      2        15             13
## 30       27       5924      1        15             14
## 31       10       6402      0        15             15
```

Now I am unsure about the indexing, as the *time bands* are aggregated $[a_{i-1}, a_i)$ and we added the extra "zero row". In that perspective, as the intervals are closed on the left and open on the right, the last row with `Follow.up = 6402` is not taken in to account. The risk set would then be 1 in the end. On the other hand, we know that the last row is censored, thus accounting the information leaves us with risk set of 0.

b)

Does the estimated survival function reach zero? If not, then why not?

Let's inspect the estimated values of the survival function

```
summary(cpa.km$surv)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##  0.1827  0.3198  0.5054  0.4739  0.5757  0.8750
```

The minimum value of the survival function is 0.1827, thus it does not reach zero.

This is because the last three patients accounted for in the estimation did not experience an event and were censored. Censoring is generally regarded as *non-informative* so it does not affect the survival probability.

```
data.frame(cpa.km$time, cpa.km$n.event, cpa.km$n.censor, cpa.km$surv) %>% tail()
```

```
##    cpa.km.time cpa.km.n.event cpa.km.n.censor cpa.km.surv
## 25        5354              1               0   0.2436656
## 26        5415              0               1   0.2436656
## 27        5639              1               0   0.1827492
## 28        5794              0               1   0.1827492
## 29        6415              0               1   0.1827492
## 30        7209              0               1   0.1827492
```

As seen from the table, after the last observed event, the estimated survival probability remains unchanged.

c)

What is the median survival time and the corresponding confidence interval?

Let's use the `print` method to see the generic output of the `survfit` object

```
print(cpa.km)
```

```
## Call: survfit(formula = Surv(time = agepres, time2 = agelast, event = dead) ~ 
##     1, data = cpa, conf.int = 0.95, conf.type = "plain", type = "kaplan-meier")
## 
## records   n.max n.start  events  median 0.95LCL 0.95UCL 
##      30      16       8      15    1849     123    5354
```

The median survival time (in days) is 1849 with the 95 % confidence interval of $[123, 5354]$ and in years

```
v <- c(1849, 123, 5354, 1849/365.25, 123/365.25, 5354/365.25)
names(v) <- c("Median", "Lower", "Upper","Median (years)", "Lower (years)", "Upper (years)")
v
```

```
##          Median          Lower          Upper Median (years)  Lower (years)
##    1849.0000000    123.0000000   5354.0000000      5.0622861      0.3367556
##  Upper (years)
##     14.6584531
```

By a visual inspection this seems to be in concordance with the survival plot above.

Q6

Would you recommend using exponential or Weibull distribution for the overall survival? Support your suggestion by using likelihood-ratio test.

We want to fit two survival models, an exponential model and a Weibull model. The exponential model a simple model with a single-parameter constant hazard rate λ as the Weibull models is a more flexible model with an additional parameter to allow for an increase or a decrease of the hazard rate over time. First we'll compare the basic characteristics of the models.

The density function $f(t;\lambda)$ for $t \geq 0$ for the exponential distribution is

$$f(t;\lambda) = \lambda \exp(-\lambda t)$$

And with a *standard* parametrization $f(t;\lambda,k)$ for the Weibull distribution is

$$f(t;\lambda,k) = \frac{k}{\lambda}\left(\frac{t}{\lambda}\right)^{k-1} \exp(-(t/\lambda)^k)$$

There are alternative parametrizations such as $f(t;\alpha,k)$ with $\alpha = \ (\frac{1}{\lambda})^k$. Let

$$\alpha = (\frac{1}{\lambda})^k = \frac{1}{\lambda^k} = \frac{1}{\lambda}\frac{1}{\lambda^{k-1}}$$

By substitution we have

$$f(t;\alpha,k) = k\alpha t^{k-1} \exp(-\alpha t^k)$$

The one in the course material, which we will be using is $f(t;\alpha,p)$. According to the `eha::weibreg` documentation the parametrization (using the letters p, α and k) is for scale parameter $\alpha = \exp(\frac{-t\beta}{k})$, where I believe $\beta = \frac{1}{\lambda}$.

Thus in our case I presume the parametrization to be $\alpha = \exp(-\frac{t}{k\lambda})$. I still have trouble *getting the hang of* this parametrization nor have I figured out what p is in relation to k, but in any case our Weibull distribution is defined as

$$f(t;\alpha,p) = p\alpha^p t^{p-1} \exp(-(\alpha t)^p)$$

As it is at all occasion clear which distribution and parametrization is discussed, I will denote the density function simply by $f(t)$. As we know the identities

$$\begin{aligned}
f(t) &= -\frac{d}{dt}S(t) = \lambda(t)S(t) \\
S(t) &= \int_t^\infty f(u)du = \exp(-\int_0^t \lambda(u)du) \\
\lambda(t) &= \frac{f(t)}{S(t)} = -\frac{d}{dt}\log(S(t))
\end{aligned}$$

the hazard rate and survival function for the exponential distribution are

$$
\begin{aligned}
f(t) &= \lambda \exp(-\lambda t) \\
S(t) &= \exp(-\lambda t) \\
\lambda(t) &= \lambda
\end{aligned}
$$

And for Weibull distribution

$$
\begin{aligned}
f(t) &= p\alpha^p t^{p-1} \exp(-(\alpha t)^p) \\
S(t) &= \exp(-(\alpha t)^p) \\
\lambda(t) &= p\alpha^p t^{p-1}
\end{aligned}
$$

We'll first fit the exponential model, for which we use the `weibreg` function. Note that the Weibull distribution is equal to the exponential distribution when the shape $p = 1$

```
cpa.exp = weibreg(Surv(time = agepres, time2 = agelast, event = dead) ~ 1,
                  data = cpa,
                  shape = 1)
```

The estimate for λ, $\hat{\lambda}$, under the exponential model is

```
lambda_exp_hat = exp(- cpa.exp$coeff[1])
lambda_exp_hat
```

```
##   log(scale)
## 0.0002883506
```

And the estimate for the survival function under the exponential model

```
t = 1:max(cpa.km$time)
surv_exp = exp(-lambda_exp_hat*t)
```

Then we'll fit the Weibull model

```
cpa.wb = weibreg(Surv(time = agepres, time2 = agelast, event = dead) ~ 1,
                  data = cpa)
```

The estimates for the Weibull distribution parameters scale and shape

```
b_hat                   = exp(cpa.wb$coeff[1])
a_hat                   = exp(cpa.wb$coeff[2])
c(b_hat,a_hat)
```

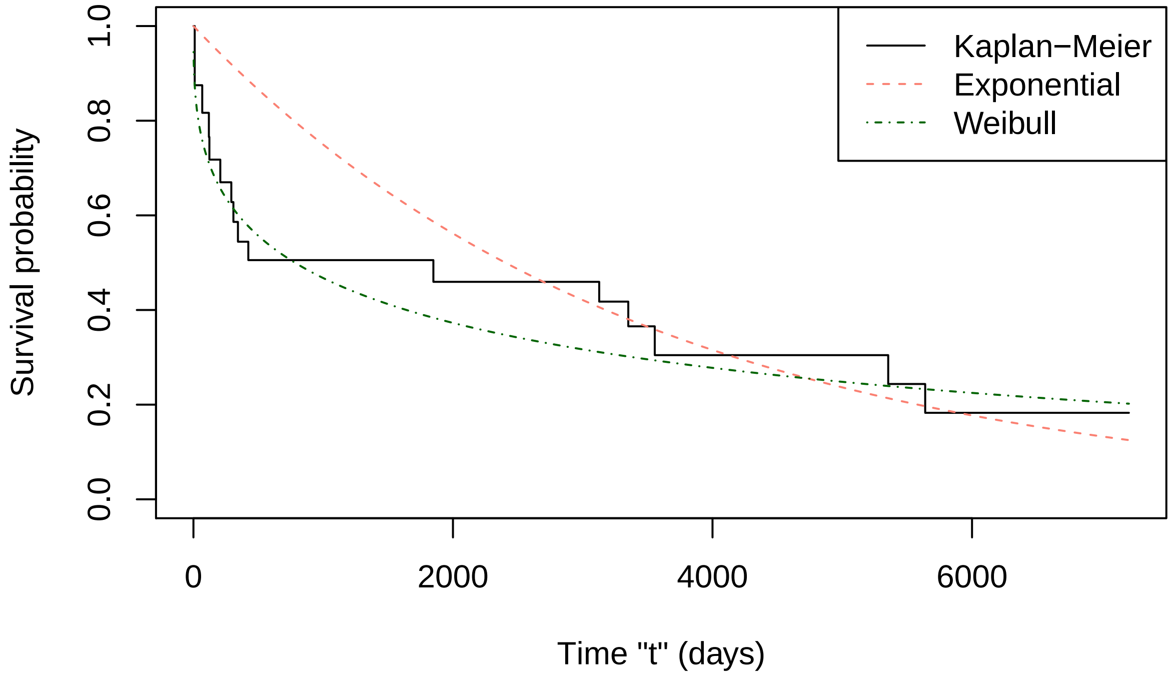
```
##   log(scale)   log(shape)
## 2071.2569994    0.3763981
```

Finally the estimate for the survival function $\exp(-(\frac{t}{a})^b)$

```
surv_wb = exp(-(t/b_hat)^a_hat)
```

Let's first compare the exponential and Weibull models visually by plotting both estimated survival functions against time. Out of pure interest let's plot them over the Kaplan-Meier estimate

```
pal <- c("black", "salmon", "darkgreen")
{
plot(cpa.km, xlab = "Time \"t\" (days)", ylab = "Survival probability",conf.int =FALSE)
lines(t, surv_exp, lty = 2,col = pal[2])
lines(t, surv_wb , lty = 4, col = pal[3])
legend("topright",legend=c("Kaplan-Meier", "Exponential","Weibull"),lty=c(1,2,4),col=pal)
}
```



So far as by any visual inspection the Weibull model seems to provide a better fit. There is compelling evidence that the hazard rate does not remain constant in time. This is hinted by the initial sharp decline of the KM survival curve, then clearly becoming less steep from approximately $t \geq 500$.

Finally let's test our claim by performing the likelihood ratio test.

Likelihood is in principle a probability function of the data conditional to the estimates of the parameters. The function is maximized over the parameter space Θ to find the Maximum Likelihood Estimate (MLE) $\hat{\theta}$

In our case we have a more complex Weibull model, for which the likelihood function we deem to be maximized over the entire Θ, and a simpler *restricted* exponential model, as in a special case of Weibull when $p = 1$, of which the likelihood function is maximized over a subset of the parameter space $\Theta_0 \subset \Theta$.

The likelihood ratio test (LRT) checks whether the restricted model, the null hypothesis, is supported by the data, i.e. is there a difference between the likelihoods of the restricted model and the unrestricted model.

Let θ_0 the MLE under the null hypothesis (restricted model) and $\hat{\theta}$ the MLE under the alternative hypothesis (unrestricted model). The likelihood ratio is

$$LR = -2\log\left(\frac{L(\theta_0)}{L(\hat{\theta})}\right)$$

The $-2\log$ of the ratio is for convenience as LR has then an asymptotic χ^2 distribution if the null hypothesis is true. For extended convenience we want to simplify the expression for the LRT

$$
\begin{aligned}
LR &= -2\log\left(\frac{L(\theta_0)}{L(\hat{\theta})}\right) \\
&= -2(\log(L(\theta_0)) - \log(L(\hat{\theta}))) \\
&= -2(l(\theta_0) - l(\hat{\theta}))
\end{aligned}
$$

In our empirical case, we can extract the log-likelihoods directly from the `weibreg` model objects

```
lr  = -2*(cpa.exp$loglik[1] - cpa.wb$loglik[1])
lr
```

```
## [1] 11.98288
```

We then want to test if the LR differs significantly from 0 with $2 - 1 = 1$ degree of freedom. As LR has an asymptotic χ^2 distribution we can apply a χ^2 test via the `pchisq` function

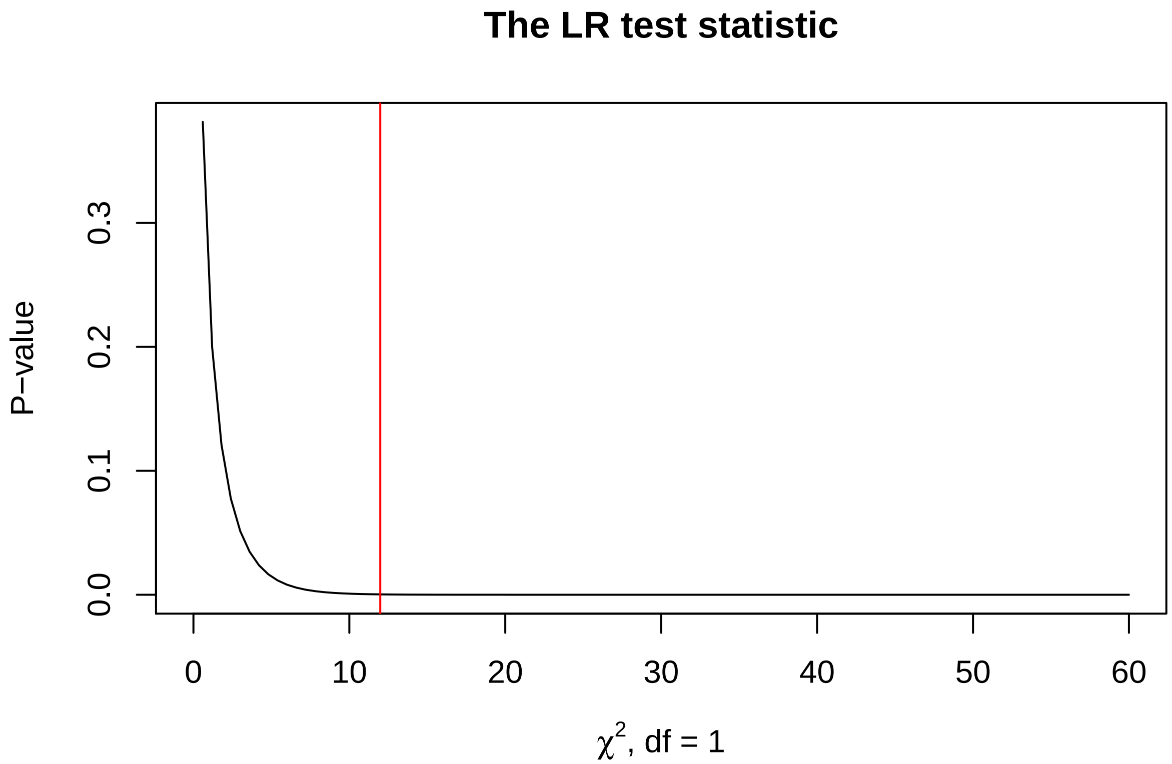
```
p_lr = 1 - pchisq(lr, df=1)

p_lr
```

```
## [1] 0.0005369152
```

The p-value represents the probability of the data given the null hypothesis that the LR is zero. In this case the data is quite critical against the null hypothesis thus we will disregard it.

We can visualize this simply by

```
{
curve( dchisq(x, df=1), main = "The LR test statistic", xlab = expression(paste(chi^2,", df = 1")), ylab
       from=0,to=60)
abline(v = lr, col = "red")
}
```



where the red line represets the test statistic.

To conclude the likelihood ratio test concords with our beliefs that the Weibull model would be a better option against the exponential model.

Q7

Perform a graphical check of proportionality for paanat. Interpret the results.

The binary `paanat` variable in the *complex pulmonary atresia* data represents the size of intrapericardial pulmonary arteries at presentation with values absent or tiny (0) and normal or near normal (1).

We want to compare the hazard rates with the two groups:

- The rate in the group with abnormal pulmonary arteries `paanat = 0`: $\lambda_0(t)$
- The rate in the group with normal pulmonary arteries `paanat = 1`: $\lambda_1(t)$

According to the proportional hazards assumption

$$\frac{\lambda_1(t)}{\lambda_0(t)} = \theta$$

for all $t \geq 0$ where θ is a constant hazard ratio.

For a binary covariate Z, such as `paanat`, we can write generally the hazard as

$$\lambda(t; \theta, Z) = \theta^Z \lambda_0(t)$$

To satisfy $\theta \geq 0$ we write the latter with $\beta = \exp(\theta)$

$$\lambda(t;\theta,Z) = \exp(\beta Z)\lambda_0(t)$$

In practice we'll stratify the the models by the aforementioned groups, thus having one model with $Z = 1$ and other with $Z = 0$. We can fit a stratified Cox proportional hazards model with

```
cpa.cox = coxph(Surv(time = agepres, time2 = agelast, event = dead) ~ strata(paanat), data = cpa)
cpa.cox
```

```
## Call:  coxph(formula = Surv(time = agepres, time2 = agelast, event = dead) ~
##     strata(paanat), data = cpa)
##
## Null model
##   log likelihood= -25.34002
##   n=30 (1 observation deleted due to missingness)
```

and create a survival object which also calculates the cumulative hazards $\Lambda(t)_{0k}$ (the integral of $\lambda(t)$ from 0 to t) for each strata k, $k = 1, 2$.

```
cpa.surv <- survfit(cpa.cox)
cpa.surv
```

```
## Call: survfit(formula = cpa.cox)
##
##          records n.max n.start events median 0.95LCL 0.95UCL
## paanat=0      20    12       6      7   3127     119      NA
## paanat=1      10     5       5      8    292     207      NA
```

NOTE

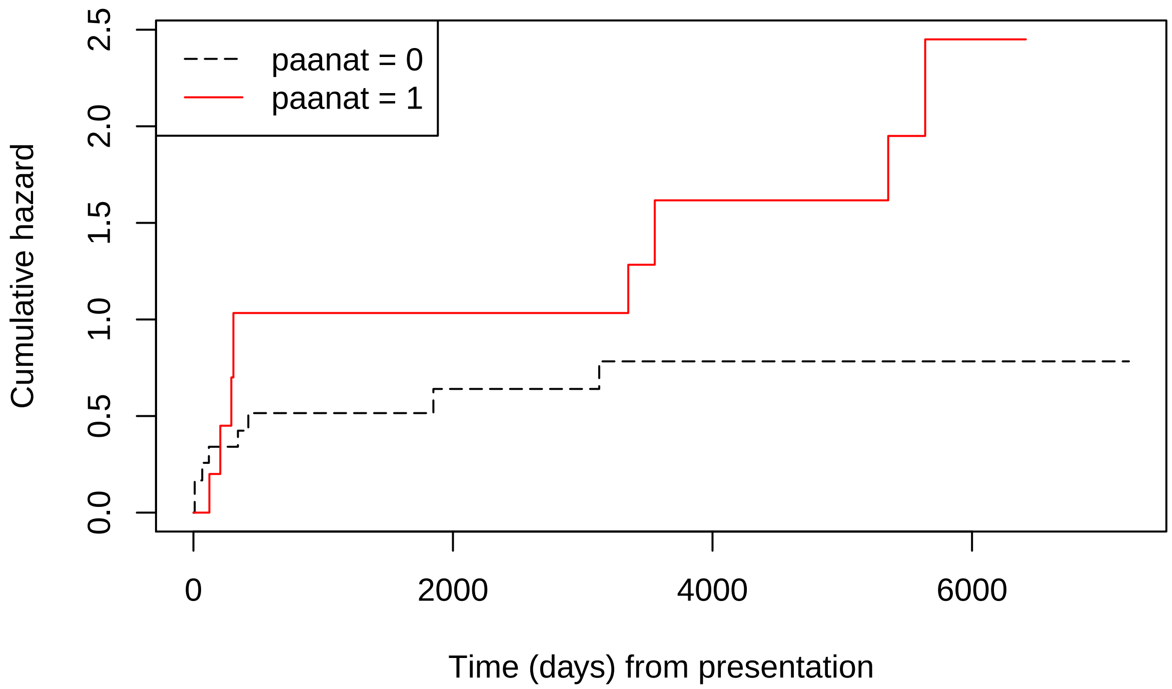
Here I am very puzzled! The values of paanat are described as 1: "normal or near normal size intrapericardial pulmonary arteries at presentation" and 0: "tiny or absent intrapericardial pulmonary arteries at presentation".

I fail to undestand why the median survival time for `paanat = 1` or "normal or near normal" is significantly less than `paanat = 0` or "tiny or absent". I assume that "normal arteries" is a good thing and "absent arteries" is bad. On `paanat=0` there are 7 events out of 20 and 8 events out of 10 with `paanat=1`.

Thus, I'll omit any intepretation of the variables themselves and raise all conlusions from the data itself.

Anyhow, we'll then plot a Nelson-Aalen plot (Cumulative hazard vs. time) of the two strata

```
{
plot(cpa.surv, fun="cumhaz", xlab="Time (days) from presentation", ylab="Cumulative hazard", lty=c(5,1)
legend("topleft",c("paanat = 0", "paanat = 1"),lty=c(5,1), col = c("black", "red"))
}
```



Graphical check implies, as the plots are not in any way parallel, that the proportionality assumption does not hold for the two strata, that is the hazard ratio is not constant for the two strata.