Survival analysis week 2

Petteri Mäntymaa

March 20, 2019

Exercise 2: Paramteric survival models and model checking

1. Properties of exponential distribution

1.1.

Let $T_1, \ldots, T_n$ be a random sample from a distribution with survival function $S(t) = exp\{\lambda t\}$. Show that the distribution of $T = nmin(T1, \ldots, Tn)$ is exponential with failure rate λ.

Note: you may prove that this result (in limit) holds even when $S(t)$ is such that $S(t) = 1\lambda t + o(t)$ as $t \to 0$ where $o(t)$ means $\frac{o(t)}{t} \to 0$ as $t \to 0$.

1.2.

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

1.3.

Show that the nth moment of the exponential distribution with falire rate λ is $E(T^n) = \frac{n!}{\lambda n}$.

ANSWER:

We have the density function $f(t) = \lambda e^{-\lambda t}$ where $t > 0$.

As the moment generating function for the exponential distribution we have

$$
\begin{aligned}
M_T(z) = E(e^{zT}) &= \int_0^\infty e^{zt} \lambda e^{-\lambda t} dt \\
&= \int_0^\infty e^{(z-\lambda)t} dt \\
&= \frac{\lambda}{z-\lambda} \Big/_0^\infty e^{(z-\lambda)t} \\
&= \frac{\lambda}{z-\lambda} \lim_{t\to\infty} \left(e^{(z-\lambda)t} - e^{(z-\lambda)0}\right) \\
&= \frac{\lambda}{z-\lambda}(0-1) \\
&= \frac{\lambda}{\lambda - z}, \ \forall \ \lambda < z
\end{aligned}
$$

With some help from my calculus book we find that we can apply the following geometric series result:

$$\frac{1}{1-x} = \sum_{n=0}^{\infty} x^n, \ \forall \ |z| < 1$$

With it we can rewrite the moment generating function as

$$M_T(z) = \frac{\lambda}{\lambda - z} = \frac{\lambda}{\lambda - \frac{z}{\lambda}} = \sum_{n=0}^{\infty} \frac{1}{\lambda^n} z^n$$

We also remember, again with a great deal of help from my calculus book, that

$$E(T^n) = M_T^{(n)}(0) = \frac{d^n}{dz^n} M_T(z), \;\; at \;\; z = 0$$

We then have to evaluate the n^{th} derivative of the moment generating function, thus we need the Maclaurin series

$$f(x) = \sum_{n=0}^{\infty} \frac{f^{(n)}(0)}{n!} x^n$$

…and the fact that $E(T^n) = M_T^{(n)}(0)$ to rewrite the moment generating function as

$$M_T(z) = \sum_{n=0}^{\infty} \frac{M_T^{(n)}(0)}{n!} z^n = \sum_{n=0}^{\infty} \frac{E(T^n)}{n!} z^n = \sum_{n=0}^{\infty} \frac{1}{\lambda^n} z^n$$

With a little bit of tidying the last two (by getting rid of coefficients and multiplying both sides with $n!$) we finally get

$$E(T^n) = \frac{n!}{\lambda^n}$$

Now this result is by no means my own invention and I gathered bits and pieces of the solution here and there but I did not utilize any parts of the solution that I did not bite and chew for a great deal first.

I do admit that many steps eluded me and I don't think I ever could have come up with them myself.

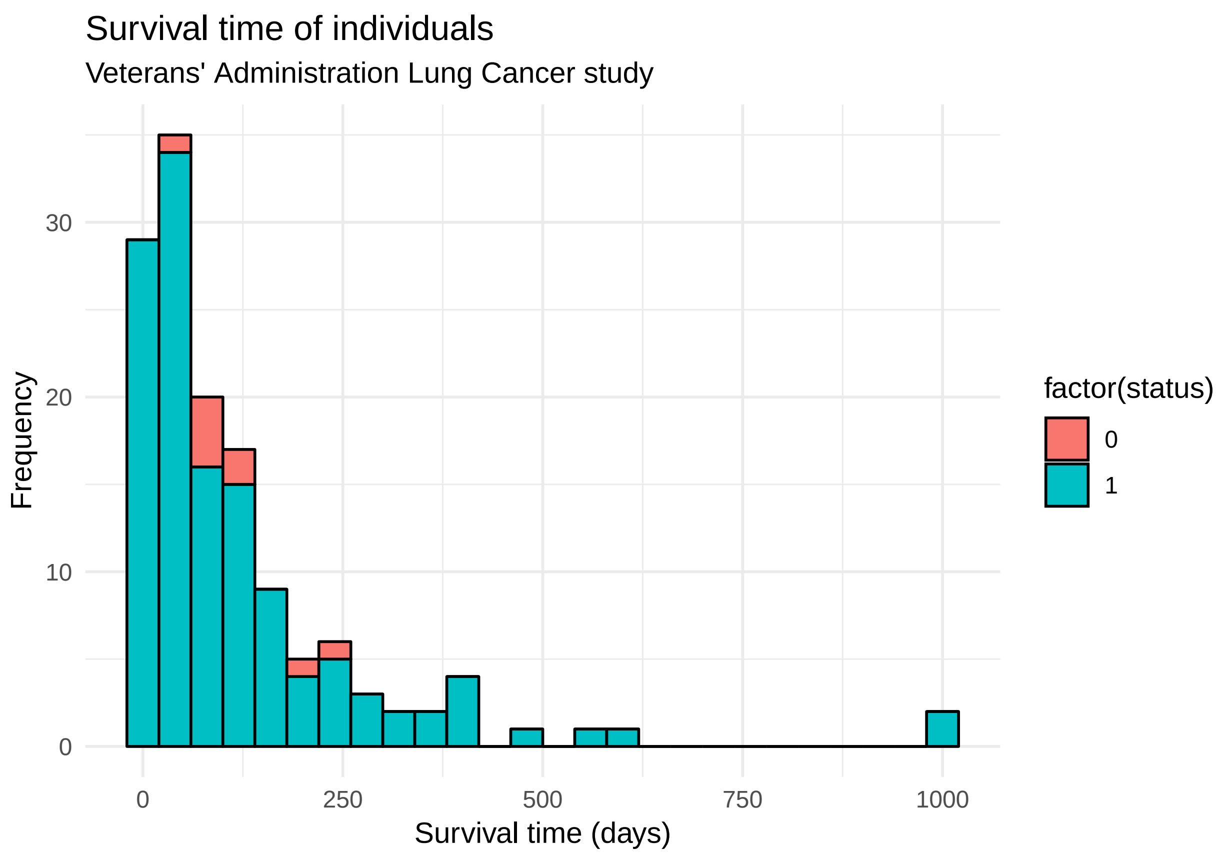
2. Fitting exponential and Weibull model to Veteran data

Load veteran data from library(survival).

2.1.

Plot a histogram of the survival times corresponding to uncensored observations (`veteran$status == 1`) as done in Exercise 1.

```
veteran %>%
  ggplot() +
  aes(x = time, fill = factor(status)) +
  geom_histogram(binwidth = 40, color = "black") +
  theme_minimal() +
  labs(title = "Survival time of individuals", subtitle = "Veterans' Administration Lung Cancer study")
  xlab("Survival time (days)") +
  ylab("Frequency")
```



Minor addition to last weeks histogram; The colour indicates censoring status, red for censored (survival time unknown) and blue for non-censored (survival time known, i.e. event occurs in the study period).

2.2.

Compare the Kaplan-Meier estimate of the survival function to

(a) Exponential distribution, and

(b) Weibull distribution

Use graphical procedure and interpret the results.

Hint: You can obtain the maximum likelihood estimates of the parameters using `weibreg()` function of `eha` package. For example, the estimate of the parameter λ in $S(t) = exp\{\lambda t\}$ is obtained using

ANSWER:

Obtain the estimate for parameter λ, create vector `t` and estimate the exponential survival probabilities for `t` as a vector `St.exp` with $S(t) = exp\{\lambda t\}$

```
veteran.exp0 <- weibreg(formula = Surv(time, status) ~ 1, data=veteran, shape=1)
lambda0 <- exp(-veteran.exp0$coeff[1])
t <- 1:1000
St.exp <- exp(-(lambda0*t))
df.exponential <- data.frame(t, St.exp)
```

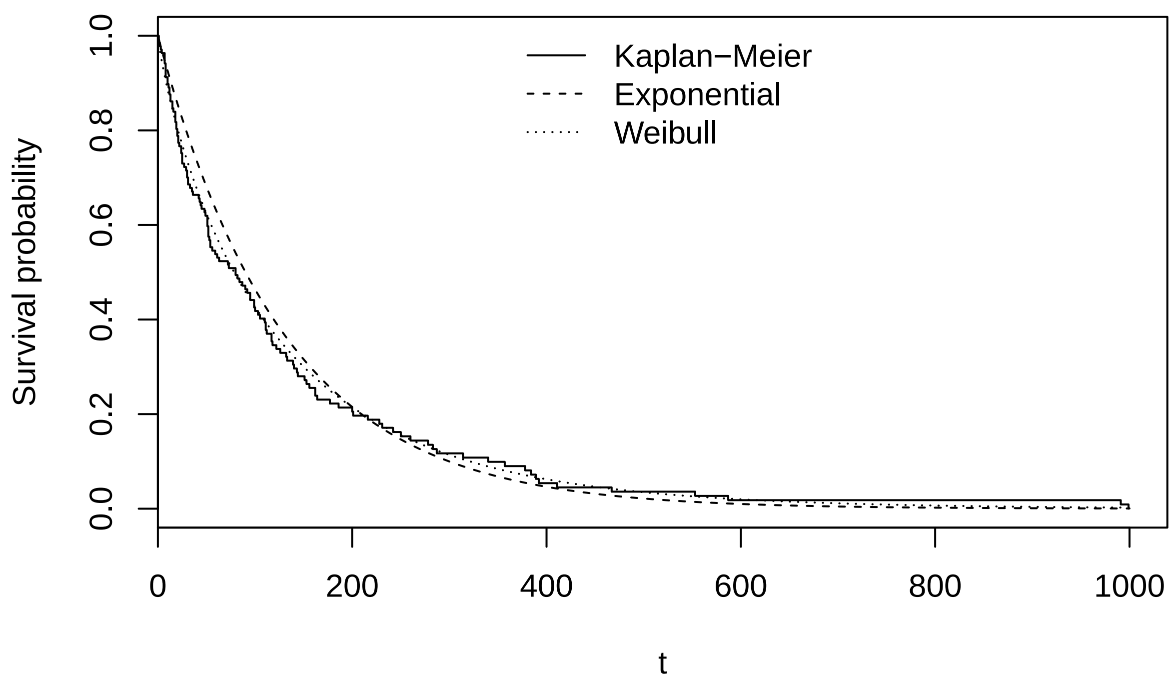
Obtain the estimates for parameters `a` and `b`, create vector `t` and estimate the weibull survival probabilities for `t` as a vector `St.w` with $S(t) = exp\{(\frac{t}{b})^a\}$

```
veteran.weibull0 <- weibreg(formula = Surv(time, status) ~ 1, data=veteran)
b <- exp(veteran.weibull0$coeff[1])
a <- exp(veteran.weibull0$coeff[2])
St.w <- exp(-(t/b)^a)
df.weibull <- data.frame(t, St.w)
```

Lastly create the Kaplan-Meier survival curves with `survfit`:

```
veteran.kaplan <- survfit(Surv(time, status) ~ 1, conf.type = "none", type = "kaplan-meier", data = vete
```

```
{
plot(veteran.kaplan, xlab = "t", ylab = "Survival probability")
lines(df.exponential$t, df.exponential$St.exp, lty = 2)
lines(df.weibull$t, df.weibull$St.w, lty = 3)
legend(x = "top", legend = c("Kaplan-Meier", "Exponential", "Weibull"), lty = c(1,2,3),bty = "n")
}
```



The exponential distribution seems to overestimate the survival probabilities approximately up to `t = 200` while the Weibull distribution seems to offer a more decent fit. Though the graphical inspection hardly offers any rigorous take on the matter, it is a important step in the process.

Model choice

2.3.

Compare the above two models with the likelihood ratio test. Interpret the result.

Hint: You can extract the log-likelihood values from the output objects of function weibreg. Use the pchisq function to calculate the p-value (tail probability).

Alternative: You can calculate the likelihood ratio by using the anova command on the output objects from the two regression models using survreg.

Obtain the log-likelihood values of the models and calculate the likelihood ratio test score. The form of the test is $LRT = -2log_e(\frac{L_1(\hat{\theta})}{L_2(\hat{\theta})})$, but our likelihoods are already log-likelihoods so we will use $-2(log_e(L_1(\hat{\theta})) - log_e(L_2(\hat{\theta})))$.

```
loglik.exp <- veteran.exp0$loglik[1]
loglik.w <- veteran.weibull0$loglik[1]

LRT <- -2*(loglik.exp - loglik.w)
LRT
```

```
## [1] 6.259992
```

with $2 - 1 = 1$ degrees of freedom.

```
pchisq(LRT, df = 1, lower.tail = F)
```

```
## [1] 0.01234947
```

The P-value indicates that there is a (quite) significant difference between the models.

3. Simulation

3.1.

Generate 100 random numbers from exponential distribution with mean 0.01 and store it in T. Before you do this, look at the help(rexp) and the parameter which it accepts.

```
options("scipen"=0, "digits"=7)
T <- rexp(100,100)
mean(T)
```

```
## [1] 0.01291214
```

3.1.1.

Plot the empirical distribution function.

3.1.2.

Esimate the rate (stored under obsrate) from the simulated data and overlay the plot of the distribution function $1 - \exp(-\text{obsrate} * \text{t})$.

3.1.3.

Overlay the plot of the true exponential distribution function.

3.1.4.

Explore the possibilities for different kinds of line and point plots. Vary the plot symbol, line type, line width, and colour. Also, try to give legend in the above graph.