Exercises and Lecture Notes, STK 4080, Autumn 2018

Version 0.02, 20-viii-2018

Nils Lid Hjort

Department of Mathematics, University of Oslo

Abstract

Exercises and Lecture Notes collected here are indeed for the Survival and Event History Analysis course STK 4080 / 9080, autumn semester 2018. The exercises will complement those given in the course book Aalen, Borgan, Gjessing, *Survival and Event History Analysis: A Process Point of View*, Springer, 2008.

1. Ancient Egyptian lifelengths

How long is a life? A unique set of lifelengths in Roman Egypt was collected by W. Spiegelberg in 1901 *(Ägyptische und griechische Eigennamen aus Mumienetiketten der römischen Kaiserzeit)* and analysed by (the very famous) Karl Pearson (1902) in the very first volume of (the very famous) Biometrika. The data set contains the age at death for 141 Egyptian mummies in the Roman period, 82 men and 59 women, dating from the last century b.C. The lifelengths vary from 1 to 96 years, and Pearson argued that these can be considered a random sample from one of the better-living classes in that society, at a time when a fairly stable and civil government was in existence (as we recall, the violent 'tax revolt' with ensuing long-lasting complications took place under Antoninus Pius later, in 139 AD). To access the data, go to `egypt-data` at the course website, reading them into your computer via

```
tt <- scan("egypt-data",skip=5)
```

Pearson did not attempt to fit any parametric models for these data, but discussed differences between the Egyptian age distribution and that of England 2000 years later. The purpose of the present exercise is to analyse aspects of the data by comparing the nonparametric survival curve (here a simplified version of the Kaplan–Meier curves, since there is no censoring; all the old Egyptians are dead) with a couple of parametric curves, in particular the Weibull.

(a) We start with the natural nonparametric estimate of the survival curve $S(t) = \Pr\{T \ge t\}$. Let the data be $t_1, \ldots, t_n$ (either the full set, or the subset for men, or that of the women). Since this is just a binomial probability, for each fixed t, we may put up the empirical survival function

$$S_{\rm emp}(t) = (1/n)\sum_{i=1}^n I\{t_i \ge t\} \quad \text{for } t > 0.$$

Show that $\mathrm{E}\, S_{\rm emp}(t) = S(t)$ and that $\mathrm{Var}\, S_{\rm emp}(t) = (1/n)S(t)\{1 - S(t)\}$.

(b) Compute the empirical survival curves, for men and for women, and display them in the same diagram, cf. Figure 0.1 below.

(c) Then consider the two-parameter Weibull model [note the Swedish pronunciation], which has a cumulative distribution of the form

$$F(t,a,b) = 1 - \exp\{-(at)^b\} \quad \text{for } t > 0,$$

with a and b positive parameters (typically unknown). (i) Find a formula for the median of the distribution. (ii) Show that the probability of surviving age t, given that one has survived up to t_0, is $\exp[-\{(at)^b - (at_0)^b\}]$, for $t > t_0$. (iii) Show that the density can be expressed as

$$f(t,a,b) = \exp\{-(at)^b\} a^b b t^{b-1} \quad \text{for } t > 0.$$

(d) Find formulae for the 0.20- and 0.80-quantiles, and set these equal to the observed 0.20- and 0.80-quantiles for the data. This yields two equations with two unknowns, which you can solve. In this fashion, find estimates $(\widetilde{a}, \widetilde{b})$ for the men and for the women.

(e) While quantile fitting is a perfectly sensible estimation method, a more generally versatile method is that of maximum likelihood (ML), which will also be used later on in the course. By definition, the ML estimates $(\widehat{a}, \widehat{b})$ are the parameter values maximising the log-likelihood function

$$\ell_n(a,b) = \sum_{i=1}^n \log f(t_i,a,b) = \sum_{i=1}^n \{-(at_i)^b + b\log a + \log b + (b-1)\log t_i\}.$$

This can be maximised numerically, as soon as you can programme the log-likelihood function. With data stored in your computer, called `tt`, try this, using R's powerful non-linear minimiser `nlm`:

```
logL <- function(para)
{
a <- para[1]
b <- para[2]
hei <- -(a*tt)^b + b*log(a) + log(b) + (b-1)*log(tt)
sum(hei)
}
# then:
minuslogL <- function(para)
{-logL(para)}
# then:
nils <- nlm(minuslogL2,c(0.20,1.00),hessian=T)
ML <- nils$estimate
```

It gives you the required ML estimates $(\widehat{a}, \widehat{b})$. Carry out this estimation scheme, for the men and the women separately.

(f) I find $(0.0270, 1.3617)$ for the men and $(0.0347, 1.5457)$ for the women. Display the two estimated Weibull survival curves, perhaps along with the two nonparametric ones, as in my Figure 0.1 here. Compute the estimatead median lifelengths, for men and for women, and comment.

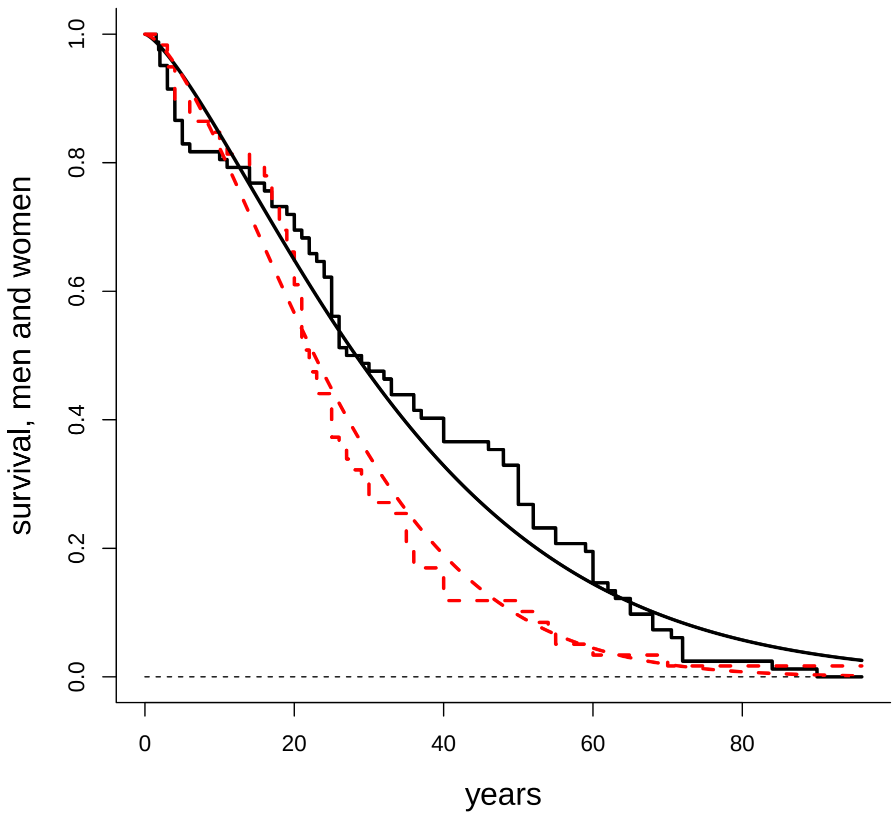

Figure 0.1: Survival curves from Roman era Egypt, two for men (black) and two for women (red). The step functions are the empirical survival curves, type Kaplan–Meier; the continuous curves are the fitted Weibull curves.

(g) Compute and display also the estimated Weibull hazard rates, for men and for women. Comment on what you find.

(h) Considerations above invite statistical testing of the hypothesis H_0 that men and women of Roman era Egypt had the same lifelength distributions. Compute and display the 90% confidence bands

$$S_{m,\mathrm{emp}}(t) \pm 1.645\,\widehat{\tau}_m(t), \quad S_{w,\mathrm{emp}}(t) \pm 1.645\,\widehat{\tau}_w(t),$$

where

$$\begin{aligned} \widehat{\tau}_m(t)^2 &= (1/n_m)S_{m,\mathrm{emp}}(t)\{1 - S_{m,\mathrm{emp}}(t)\}, \\ \widehat{\tau}_w(t)^2 &= (1/n_w)S_{w,\mathrm{emp}}(t)\{1 - S_{w,\mathrm{emp}}(t)\}, \end{aligned}$$

the estimated variances. (We shall learn formal tests along such lines in the course.)

(i) Above I've forced you through the loops of things for one particular parametric model, namely the Weibull. Now do all these things for the Gamma(a, b) model too, with density $\{b^a/\Gamma(a)\}t^{a-1}\exp(-bt)$. Part of the point here is that this does not imply a doubling of your work efforts; you may edit your computer programmes, at low work cost, to accommodate other parametric models, once you've been through one of them. The Weibull does a slightly better job than the Gamma, it turns out.

"Either man is constitutionally fitter to survive to-day [than two thousand years ago], or he is mentally fitter, i.e. better able to organise his civic surroundings. Both conclusions point perfectly definitely to an evolutionary progress." – *Karl Pearson*, 1902.

2. Maximum likelihood estimation with censored data

well:

3. Gamma Hazards

well:

References

Aalen, O.O., Borgan, Ø., and Gjessing, H.K. (2008). *Survival and Event History Analysis: A Process Point of View.* Springer.

Andersen, P.K., Borgan, Ø., Gill, R., and Keiding, N. (1993). *Statistical Models Based on Counting Processes.* Springer.

Billingsley, P. (1968). *Convergence of Probability Measures.* Wiley, New York.

Claeskens, G. and Hjort, N.L. (2008). *Model Selection and Model Averaging.* Cambridge University Press, Cambridge.

Cunen, C., Hjort, N.L., and Nygård, H. (2018). Statistical sightings of better angels. Submitted for publication.

Ghosal, S. and van der Vaart, A. (2017). *Fundamentals of Nonparametric Bayesian Inference.* Cambridge University Press, Cambridge.

Hjort, N.L. (1985). Discussion contribution to P.K. Andersen and Ø. Borgan's 'Counting process models for life history data: A review'. *Scandinavian Journal of Statistics* **12**, xx–xx.

Hjort, N.L. (1985). An informative Bayesian bootstrap. Technical Report, Department of Statistics, Stanford University.

Hjort, N.L. (1986). Discussion contribution to P. Diaconis and D. Freedman's paper 'On the consistency of Bayes estimators'. *Annals of Statistics* **14**, 49–55.

Hjort, N.L. (1990). Nonparametric Bayes estimators based on Beta processes in models for life history data. *Annals of Statistics* **18**, 1259–1294.

Hjort, N.L. (1991). Bayesian and empirical Bayesian bootstrapping. Statistical Research Report, Department of Mathematics, University of Oslo.

Hjort, N.L. (2003). Topics in nonparametric Bayesian statistics [with discussion]. In *Highly Structured Stochastic Systems* (eds. P.J. Green, N.L. Hjort, S. Richardson). Oxford University Press, Oxford.

Hjort, N.L. (2018). Towards a More Peaceful World [Insert '!' or '?' Here]. FocuStat Blog Post.

Hjort, N.L., Holmes, C.C., Müller, P., and Walker, S.G. (2010). *Bayesian Nonparametrics.* Cambridge University Press, Cambridge.

Hjort, N.L. and Petrone, S. Nonparametric quantile inference using Dirichlet processes. In *Festschrift for Kjell Doksum* (ed. V. Nair).

Pearson, K. (1902). On the change in expectation of life in man during a period of circa 2000 years. *Biometrika* **1**, 261–264.

Schweder, T. and Hjort, N.L. (2016). *Confidence, Likelihood, Probability.* Cambridge University Press, Cambridge.