STK 4080-9080 Autumn 2018
The Oblig

This is The Oblig, the mandatory assigment, for STK 4080-9080, Survival Analysis & Event History Analysis, Autumn 2018. It is made available at the course website Tuesday October 30, and the submission deadline is Tuesday November 13, 13:59, via the Devilry system (`devilry.ifi.uio.no`). Reports may be written in nynorsk, bokmål, riksmål, English or Latin, should be text-processed (for instance with TeX or LaTeX), and must be submitted as a single pdf file. The submission must contain your name, the course, and assignment number.

The Oblig set contains two exercises and comprises five pages (in addition to the present introduction page, 'page 0').

It is expected that you give a clear presentation with all necessary explanations, but write concisely (in der Beschränkung zeigt sich erst der Meister; brevity is the soul of wit; краткость – сестра таланта). Remember to include all relevant plots and figures. These should preferably be placed inside the text, close to the relevant subquestion.

For a few of the questions setting up an appropriate computer programme might be part of your solution. The code ought to be handed in along with the rest of the written assignment; you might place the code in an appendix.

Students who fail the assignment, but have made a genuine effort at solving the exercises, are given a second attempt at revising their answers. All aids, including collaboration, are allowed, but the submission must be written by you and reflect your understanding of the subject. If we doubt that you have understood the content you have handed in, we may request that you give an oral account.

Application for postponed delivery: If you need to apply for a postponement of the submission deadline due to illness or other reasons, you have to contact the Student Administration at the Department of Mathematics (email: `studieinfo@math.uio.no`) well before the deadline.

The mandatory assignment in this course must be approved in the same semester, before you are allowed to take the final examination.

Complete guidelines about delivery of mandatory assignments can be found here:

`www.uio.no/english/studies/admin/compulsory-activities/mn-math-mandatory.html`

Enjoy [imperative pluralis].

Nils Lid Hjort

[Footnote: I've been inspired by Céline Cunen's Revolutionary Oblig for her course STK 4021-9021 on Bayesian analysis (also Autumn 2018, given two weeks before the present Oblig). For Exercise 1 I'm using the dataset Cunen has managed to construct, but ask different questions.]

1. Presidents of the First Republic

LIBERTÉ, ÉGALITÉ, FRATERNITÉ is the famous motto of the French Republic. At the time of its origination, during the French revolution, the motto often came with a darker addition, – OU LA MORT, and that is precisely the aspect we will examine in this exercise.

Revolutionary days are a time for experimentation, and a number of new systems were tested out during the First Republic (1792–1804), for example a new calendar system, several new state religions (among others the Cult of Reason and the Cult of the Supreme Being), and, naturally, new political systems. One such system was the National Convention (of 749 elected members), whose president could then be considered France's legitimate Head of State in this period. The presidents were elected for 14 day terms, and this gives us an interesting dataset of $n = 73$ different French presidents for the full National Convention period (September 1792 to November 1795)*. You'll find the dataset `french-data.csv` on the course website, with 73 rows and several columns with various information for the presidents (including their names). You may read the data into your computer, along with suitable names for its columns, using something like this:

```
french <- read.csv("french-data.csv")
birth <- french[ ,3]
death <- french[ ,4]
presistart <- french[ ,5]
presiend <- french[ ,6]
vv <- french[ ,7] # violent death or not
gironde <- french[ ,8]
vip <- french[ ,9]
```

These columns provide information about the lives of these presidents, with his birth and death times (with year, month, day translated into the continuous calendar scale), then the start time and end time of his presidency (similarly transcribed). The three last columns are `vv`, the dramatic indicator of having had a violent death or not; `gironde`, with 1 for having belonged to the political faction la Gironde of the First Republic, and 0 for having belonged to one of the others, like Montagne or Marais; and finally what I call `vip` here is a proxy score for having been a very important person, defined here, by Cunen, as the number of languages in which there is a wikipedia article about the president in question.

In the following, we take an interest the time t_i it took president i to die, from the end of his presidency, for the 73 presidents. To analyse such data we may operate with two viewpoints, so to speak. Perspective A is that 'a life is a life and you die when you die', and since we know the t_i for each president there is no statistical censoring. Perspective B is different, and holds that the life of a guillotined man 'should' have been longer, so we then consider the imagined what-if lifetime t_i^* to have been unpleasantly censored. In

* Four presidents were elected for two (non-consecutive) terms, but for these presidents we have only provided the dates for the last term.

yet other words, with perspective B, the survival data are of the form (t_i, δ_i), with $\delta_i = 0$ for those having met a violent death and $\delta_i = 1$ for those lucky enough not to have been killed.

(a) With viewpoint A, just described, compute and display the Nelson–Aalen estimator for the implied cumulative hazard function. Use a couple of lines to explain what the estimator is telling us here. What is the estimated median time, say $\widehat{m}_A$, from end of presidency to death?

(b) We now leave viewpoint A and for the rest of this exercise operate under perspective B. Compute and display the Nelson–Aalen estimator for the underlying cumulative hazard rate, along with an approximate 90% pointwise confidence band. Compute and display also the Kaplan–Meier estimator, and read off the estimated median time from end of presidency to death, say $\widehat{m}_B$, from this plot. Comment on the two median estimates $\widehat{m}_A$ and $\widehat{m}_B$.

(c) Find an approximate 90% confidence interval for the median time m_B from end of presidency to death.

(d) We shall now fit a parametric Gompertz model to the survival data (t_i, δ_i), with hazard rate function of the form

$$\alpha(s) = \theta \exp(\gamma s) \quad \text{for } s > 0.$$

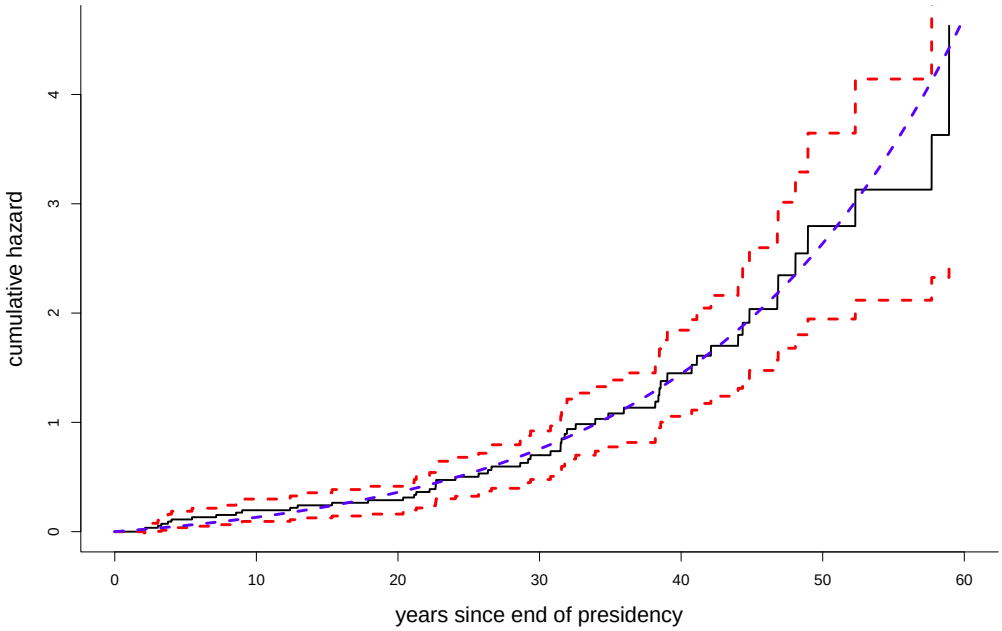
Let N_i and Y_i be the counting process and at-risk process associated with president i. With τ an upper finite limit for the time observation window in question, like 75 years, show that the log-likelihood function may be expressed as

$$\begin{aligned} \ell_n(\theta, \gamma) &= \sum_{i=1}^n \int_0^\tau \{\log \alpha(s, \theta, \gamma)\, \mathrm{d}N_i(s) - Y_i(s)\alpha(s, \theta, \gamma)\, \mathrm{d}s\} \\ &= \sum_{i=1}^n (\log \theta + \gamma t_i)\delta_i - \sum_{i=1}^n \theta \frac{\exp(\gamma t_i) - 1}{\gamma}. \end{aligned}$$

(e) Fit the two-parameter Gompertz model above, by numerically finding the maximum likelihood estimates $(\widehat{\theta}, \widehat{\gamma})$ in question (I find $(0.0099, 0.0551)$). You may e.g. use the R algorithm `nlm` to minimise the negative log-likelihood function. Compute also approximate standard errors (estimated standard deviations) for the two parameter estimates.

(f) Attempt to produce a version of the figure below, and comment on what might be learned from that figure.

(g) Using the Gompertz model, estimate the probability p_B that a newly retired president will live for at least twenty more years, supposing and praying he is not beheaded. Also give a 90% confidence interval for this p_B. How different is your estimate of p_A, the corresponding probability under perspective A?



Nelson–Aalen plot for the cumulative hazard rate for the time from end of presidency to death, with 90% confidence band, along with estimated cumulative hazard rate based on the Gompertz model (blue dashed curve).

(h) Now push the two covariates `gironde` and `vip` into the bargain. Attempt to fit the four-parameter Gompertz hazard rate regression model

$$\alpha_i(s) = \theta \exp(\gamma s) \exp(\beta_1 \,\text{gironde}_i + \beta_2 \,\text{vip}_i) \quad \text{for } i = 1, \ldots, n.$$

Use theory from the course material to give standard errors for the four parameter estimates, and test the two hypotheses $\beta_1 = 0$ and $\beta_2 = 0$. Summarise your findings.

(i) Carry out standard analysis using the Cox's proportional hazards regression model, as an alternative to what you're asked for in question (g). Comment on your findings, regarding the possible influence of la Gironde or of `vip` on the life-time distributions.

(j) If you see other statistical questions of relevance to the study of post-presidential life for the First Republic, and which can be worked with using this or perhaps a suitably extended dataset, tell us.

2. Estimating and assessing a relative change function

Consider survival data of the usual form (t_i, δ_i), with t_i the perhaps censored survival time and δ_i the indicator for non-censoring, for individuals $i = 1, \ldots, n$. We assume these individuals form a representative sample from a certain subpopulation with an underlying common hazard rate function $\alpha(s)$. One wishes to compare this subpopulation with a broader background population, where the hazard rate has been estimated and assessed earlier, with good precision, and here taken as a known $\alpha_0(s)$.

To compare the new group's hazard $\alpha(s)$ with the known background hazard $\alpha_0(s)$, write

$$\alpha(s) = \alpha_0(s)r(s) \quad \text{for } s \geq 0,$$

where the object is to estimate the relative change function $r(s)$, and perhaps to see where $r(s)$ might be significantly larger than 1.

(a) Consider the cumulative relative change function $R(t) = \int_0^t r(s)\,\mathrm{d}s$. With the usual setup, with counting process N and at-risk process Y, show that

$$M(t) = N(t) - \int_0^t Y(s)\alpha_0(s)r(s)\,\mathrm{d}s \quad \text{for } t \geq 0$$

is a martingale.

(b) Show also that the variance process for this martingale can be expressed as

$$\langle M, M\rangle(t) = \int_0^t Y(s)\alpha_0(s)r(s)\,\mathrm{d}s \quad \text{for } t \geq 0.$$

(c) Argue that

$$\widehat{R}(t) = \int_0^t \frac{\mathrm{d}N(s)}{Y(s)\alpha_0(s)} \quad \text{for } t \geq 0$$

is a natural estimator for $R(t)$. Give a formula for this estimator expressed as a finite sum. Explain why the $\widehat{R}(t)$ is a generalisation of the Nelson–Aalen estimator.

(d) Define first $R^*(t) = \int_0^t J(s)r(s)\,\mathrm{d}s$, where $J(s) = I\{Y(s) \geq 1\}$. Show that

$$\widehat{R}(t) - R^*(t) = \int_0^t \frac{\mathrm{d}M(s)}{Y(s)\alpha_0(s)}.$$

(e) Give a formula for the variance of $\widehat{R}(t) - R^*(t)$, along with a natural estimator for this variance.

(f) Attempt to give a clear limit distribution result of the type

$$\sqrt{n}\{\widehat{R}(t) - R(t)\} \to_d W(t),$$

where this limit is a zero-mean Gaußian process with a certain variance $v(t) = \mathrm{Var}\, W(t)$. Give also an estimator for this $v(t)$.

(g) How can you test the null hypothesis that the subpopulation under study here has the same hazard rate as the background population?

(h) In order to illustrate the use of this $\widehat{R}(t)$ machinery you are to simulate datapoints t_i from the model with $\alpha(s) = \alpha_0(s)r(s)$, with

$$\alpha_0(s) = \theta_0 s^{\gamma_0}, \quad r(s) = s^{\kappa_0}, \quad \text{so that } \alpha(s) = \theta_0 s^{\gamma_0+\kappa_0}.$$

So the null model is a Weibull, with parameters (θ_0, γ_0), but the real model is a different Weibull, with parameters $(\theta_0, \gamma_0 + \kappa_0)$. Show first that if you start with simulated v_i from the unit exponential, then the transformed variables

$$t_i = \{(\gamma+1)v_i/\theta\}^{1/(\gamma+1)}$$

are coming from the Weibull with parameters (θ, γ), i.e. with the desired hazard rate θs^γ.

(i) Using the simulation recipe above, simulate $n = 250$ datapoints from the Weibull with parameters $(\theta_0, \gamma_0 + \kappa_0)$, as above, with $\theta_0 = 1.00$, $\gamma_0 = 0.50$, $\kappa_0 = 0.25$. For simplicity take these as observed data, without complicating the picture with censoring mechanisms. Compute and display $\widehat{R}(t)$, along with an approximate 95% pointwise confidence band, and with $\alpha_0(s) = \theta_0 s^{\gamma_0}$ as the null model. Construct also a plot for testing whether the data are coming from the null model, i.e. for testing whether $r(s) = 1$. Comment on your findings.

One of the presidents of our dataset is particularly conspicuous by having wikipedia articles in a very large number of languages, but escaping execution. He survived to paint the next rulers of France. Here is Napoleon Crossing the Alps, *by Jacques-Louis David.*

It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness, it was the epoch of belief, it was the epoch of incredulity, it was the season of Light, it was the season of Darkness, it was the spring of hope, it was the winter of despair, we had everything before us, we had nothing before us, we were all going direct to Heaven, we were all going direct the other way – in short, the period was so far like the present period, that some of its noisiest authorities insisted on its being received, for good or for evil, in the superlative degree of comparison only.