Sample questions 2019

Statistical Lifetime Models
Lecturer: David Steinsaltz

April 18, 2019

Do not turn this page until you are told that you may do so

Note: The first two questions are from the 2018 exam. The last question has not appeared on an exam for this course.

1. (a) [10 marks]

(i) State (without derivation) the *Kaplan–Meier estimator* and the *Nelson–Aalen estimator* for the survival function of a population subject to left truncation and right censoring. Define clearly the notation that you use.

(ii) State (without proof) an estimator for the variance of the Kaplan–Meier cumulative hazard $-\log \widehat{S}(t)$ for fixed t, and explain how it may be used to derive a 95% confidence interval for $S(t)$.

(iii) Following the procedure above, we draw an upper survival function connecting all the upper bounds of the 95% confidence intervals for $S(t)$, and a lower survival function connecting the lower bounds. Should you be about 95% confident that the true survival function $S(t)$ lies entirely between the upper and lower curves? Explain why or why not.

(iv) The *Fleming–Harrington test* is based on the test statistic

$$\frac{\sum_{j=1}^{m} W(t_j)\big(d_{j1} - n_{j1}\frac{d_j}{n_j}\big)}{\sqrt{\sum_{j=1}^{m} W(t_j)^2 \frac{n_{j1}n_{j2}(n_j-d_j)d_j}{n_j^2(n_j-1)}}}.$$

Explain what the terms here mean, what the weight function W is, and how the statistic is used to compare survival distributions between two populations.

(v) The Fleming–Harrington weight function has two parameters, p and q. Explain why you might in one setting prefer $p = 1$ and $q = 0$, and in another setting might prefer $p = 0$ and $q = 1$.

(b) [15 marks] Researchers studying a novel pathogen are interested to discover how long it takes for individuals who were exposed to develop detectable antibodies. They recruit 20 subjects who are known to have been exposed at a definite time, and test them weekly for antibodies. It is believed that those who have not developed antibodies after 24 weeks will never develop them.

Below are the data they collected. "Recruit" is the number of weeks after exposure that they were recruited into the study; "Time" is the number of weeks after exposure when antibodies were found (if "event" is 1) or left the study (if "event" is 0).

Patient ID	Recruit	Time	event	Patient ID	Recruit	Time	event
1	1	16	0	11	9	15	1
2	1	8	1	12	10	15	1
3	4	8	1	13	10	18	1
4	3	9	1	14	12	18	1
5	6	9	1	15	2	24	0
6	2	8	0	16	8	24	0
7	2	8	0	17	9	24	0
8	4	8	0	18	9	24	0
9	6	8	0	19	12	24	0
10	6	8	0	20	12	24	0

(i) Some people were initially recruited for the study, but were excluded because they already had antibodies. What do we call this phenomenon?

(ii) Suppose the people who already had antibodies when they were recruited were included, recording the number of weeks since exposure. What would we call this kind of observation?

(iii) Compute the Kaplan–Meier estimator for the survival function $S(t)$, the probability that there are no antibodies after t weeks.

(iv) Subjects 6 through 10 were removed from the study at week 8 because they caught a cold that was going around, and it was thought that their blood tests would be unreliable. Later the question is raised whether this censoring was non-informative. Blood samples for these five subjects are reanalysed, and it is discovered that four of them never developed the antibodies, while the fifth developed antibodies at time $t = 9$. Define *non-informative censoring*, and carry out a statistical test to decide whether the censoring of these five subjects could have been non-informative.

2. (a) [12 marks] Suppose we have a relative risk model where individual i, with one-dimensional covariates given by $x_i(t)$ at time t, has hazard rate $r(\beta, x_i(t))h_0(t)$ at time t, where $x_i(t)$ is observed, h_0 is an unknown function, and $\beta \in \mathbb{R}$ is unknown.

We observe for each of n individuals right-censored data (T_i, δ_i), where δ_i is the indicator of an uncensored observation. We also observe the covariate processes for individual i completely up to T_i.

(i) Write down the formula for the partial likelihood, and explain how it may be used to estimate the parameters β. Define all variables used in your expression. Allow for the possibility that some individuals have identical event times.

(ii) Suppose we have calculated an estimator $\hat{\beta}$ for the parameter. Write down Breslow's estimator $\widehat{H}_0(t)$ for the baseline cumulative hazard H_0, now assuming that the event times are all distinct.

(iii) Suppose we use the Cox relative-risk function $r(\beta, x) = \mathrm{e}^{\beta x}$. The estimated parameter is $\hat{\beta} = -0.693$, with $\exp(\hat{\beta}) = 0.500$. We have computed Breslow's estimator for the baseline cumulative hazard. The values at some particular times are listed in the following table:

t	$\widehat{H}_0(t)$
0	0
1	0.25
2	0.60
3	1.00
4	2.00

An individual has covariates observed to be $x_i(t) = 1$ for $0 \le t \le 1$, $x_i(t) = 2$ for $1 \le t < 3$, and $x_i(t) = 0$ for $t \ge 3$. No events occur exactly at times $0, 1, 2, 3, 4$. Compute an estimate for the cumulative hazard specific to this individual.

(iv) Explain how to compute *Cox–Snell residuals*, and describe a graphical plot that you could make to test whether the data are well fit by this model.

(v) Suppose we suspect that the effect of x is not constant over time. Describe a plot that could be used to detect this, and explain how it would be interpreted to show that the effect of x increases over time.

(b) [13 marks] Below is a portion of a cohort life table for people born in England and Wales in 1841:

Age	q_x	l_x	d_x	E_x^c	e_x°
0	0.147	100000	14671	89814	41.6
1	0.064	85329	5422	82461	47.7
...	...	...	...	...	...
60	0.029	38029	1110	37543	15.1
...	...	...	...	...	...

(The column labelled E_x^c gives the number of years lived at age x for the nominal l_x individuals who were alive on their x-th birthday.)

(i) What is the *radix* of this life table?

(ii) Of those children who died in their first year, what was the average length of life?

(iii) Compute ${}_2q_0$.

(iv) The last column gives expected residual lifetime at age x. Calculate e_2°.

(v) Of those born in 1841 who reached their second birthday, what fraction lived to celebrate a birthday in 1901?

(vi) Suppose the mortality rate was constant in the first month, and then a different constant in the next 11 months. Estimate these two rates.

3. (a) [7 marks] A medical device manufacturer is interested in the factors that affect the lifetimes of replacement hips. They study 1000 patients who have had their prostheses implanted over the past 10 years. They collect data on whether the device needed to be replaced, and if so when; age; weight; sex; one of three different diagnosis classes as reason for the replacement.

(i) Describe how these data would be used to specify a Cox proportional hazards model, and how the results would be interpreted to decide whether men or women tend to have their replacement hips last longer.

(ii) Suppose they suspect that age might have a nonlinear effect on the hazard rate. Describe a plot that they might make to detect this, and how the plot might be used to propose a more appropriate form for the effect. (Define your terms explicitly.)

(iii) Suppose they are interested to find individuals whose prostheses failed much earlier than you would have expected, on the basis of the factors that they have already considered. What calculation would they do to find them? (Define your terms explicitly.)

(b) [10 marks] Suppose we have a relative risk model where individual i, with covariates given by a p-dimensional vector $x_i(t)$ at time t has hazard rate $r(\beta, x_i(t))\alpha_0(t)$ at time t, where α_0 is an unknown function, and $\beta \in \mathbb{R}^p$ is unknown.

We observe for each of n individuals right-censored data (T_i, δ_i), where δ_i is the indicator of an uncensored observation. There are no ties among the times T_i. We also observe the covariate processes for individual i completely up to T_i.

(i) Write down the formula for the partial likelihood, and explain how it may be used to estimate the parameters β. Define all variables used in your expression.

(ii) Write down Breslow's estimator for the baseline cumulative hazard A_0.

(iii) Describe a hypothesis test for the null hypothesis $H_0 : \|\beta\| = 0$ (that is, all parameters β_j are 0).

(iv) Suppose now that we have an individual with known covariate values $x_0(t)$ at all times t. Write an expression for estimating the individual survival function for this individual.

(c) [8 marks] A survival experiment is conducted with n individuals. Each individual i has a known fixed covariate x_i, and the covariates are all distinct. It is believed that the mortality rate for individual i at time t is of the form $\alpha_i(t) = \beta_0(t) + \beta_1(t)x_i$, where β_0 and β_1 are both unknown. Let $B_j^*(t) = \int_0^t J(s)\beta_j(s)ds$, where $J(s)$ is the indicator of the number of subjects remaining under observation being at least 2. There is independent right censoring. $\mathbf{B}^*(t)$ is the vector $\begin{pmatrix} B_0^*(t) \\ B_1^*(t) \end{pmatrix}$, and $\mathbf{N}(t)$ is an n-dimensional vector whose i-th component is 1 if subject i has died before time t, and 0 if i is still alive or censored. The event times are all distinct.

Let $R(t)$ be the set of individuals still under observation at time t, and let

$$\mu_k(t) = \frac{1}{\#R(t)} \sum_{i \in R(t)} x_i^k.$$

Let $t_1 < t_2 < \cdots < t_m$ be the times of observed events, and let i_j be the unique individual who died at time t_j.

(i) Explain why

$$\sum_{t_j \le t} \mathbf{X}^-(t_j)_{\cdot i_j}$$

is an unbiased estimator for $\mathbf{B}^*(t)$ for appropriate choice of $\mathbf{X}^-(t)$. Explain what i_j is, and give a formula for computing $\mathbf{X}^-(t)$.

(ii) State one advantage of additive hazards regression over relative risk regression. (This should be a general advantage, not the fact that in a particular case additive hazards may be better suited to the data.)