Survival Analysis - Winter 2019

Assignment 1

The assignment is due January 30 before the class. The dataset needed in Questions 4 can be found from `http://individual.utoronto.ca/osaarela/finrisk82.csv` and the R code for writing the input data files for the example model from `http://individual.utoronto.ca/osaarela/poissonreg.r`.

1. Hoem (1987, *International Statistical Review* 55) considered the association between marital status and mortality of young German men using the below dataset.

Table 1: Number of deaths and person years of German males by age and marital status

Age	Deaths			Person Years		
	Single	Married	Total	Single	Married	Total
22	433	24	457	91,444	8,556	100,000
23	412	36	448	86,835	12,708	99,543
24	337	66	439	75,892	23,203	99,095
25	331	102	433	63,241	35,415	98,656
26	287	138	425	52,023	46,207	98,230
27	242	171	413	42,123	55,675	97,798
28	215	185	400	36,915	60,470	97,385
29	192	200	392	32,215	64,770	96,985
Total	2,485	922	3,407	480,688	307,004	787,692

   (a) Specify a model for the mortality rate, including an intercept term, age group effects, and marital status effect. Assume that the marital status effect is proportional over age.

   (b) Write the likelihood function for the parameters in the model you specified in (a).

   (c) Show how the regression coefficient parameter for marital status can be interpreted in terms of a rate ratio.

   (d) Fit the model and interpret the results.

   (e) Calculate the expected number of events in each age/marital status category. Compare the expected numbers to observed event counts to assess the overall model fit. What statistical test can you use for this?

   (f) Fit also a model that allows for interaction between marital status and age. What can you say about the model fit now?

2. Type I censoring refers to censoring that is predetermined so that anyone who has not yet experienced the outcome event by the end of follow-up period at time τ is censored at τ. Consider a type I censored sample assumed to be generated by a constant

Olli Saarela, Dalla Lana School of Public Health, 155 College Street, 6th floor, Toronto, Ontario M5T 3M7, Canada. E-mail: `olli.saarela@utoronto.ca`

hazard rate λ, with censoring time τ common to all n individuals followed up. The observed data are realizations of (T_i, E_i) for $i = 1, \ldots, n$, where $T_i \equiv \min\{\tilde{T}_i, \tau\}$, and $E_i \equiv \mathbf{1}_{\{T_i = \tilde{T}_i\}}$.

(a) Derive the likelihood expression for λ assuming that the latent event times $\tilde{T}_i$ are exponentially distributed.

(b) Show that the same expression as in (a) would be obtained by assuming that $\sum_{i=1}^{n} e_i$ is Poisson distributed (conditional on $\sum_{i=1}^{n} t_i$).

(c) Suppose that we don't observe T_i but only E_i for $i = 1, \ldots, n$. Show that $\sum_{i=1}^{n} e_i$ is binomially distributed, and write the likelihood expression for λ.

(d) Using the likelihood expression obtained in (a)-(b), derive the maximum likelihood estimator for λ, and a variance estimator for this.

(e) **Bonus question, extra marks are possible.** Derive the maximum likelihood estimator for λ using the likelihood expression in (c), and a variance estimator for this. Compare the efficiency of the two maximum likelihood estimators.

3. Homogeneous Poisson process $N(t)$ counts events occurring in a time interval and is characterized by $N(0) = 0$ and $N(t + \tau) - N(t) \sim \text{Poisson}(\lambda\tau)$, where τ is the length of the interval.

(a) Show that the interarrival times to next event are independent and exponentially distributed random variables.

(b) A random variable X is said to be memoryless if

$$P(X > s + t \mid X > t) = P(X > s) \quad \forall \quad s, t \geq 0.$$

Show that this property applies for the interarrival times if they are exponentially distributed.

(c) **Bonus question, extra marks are possible.** Show that the property in (b) applies for the interarrival times only if they are exponentially distributed.

(d) Show that the conditional distribution of $N(s) = k$ given $N(t) = n$, where $s < t$ and $k \leq n$, is binomial.

(e) Suppose that patients arrive at an emergency room at the rate of $\lambda = 50$ per day.

i. What is the expected time (in hours) for 10 new patients to arrive.

ii. What is the probability that an hour goes by without a new patient arriving?

4. The trend for total mortality in the example dataset used for demonstrating piecewise constant hazard models was decreasing over calendar time. We might be interested whether this decrease is due in particular to decrease in coronary heart disease (CHD) mortality.

(a) Using the provided dataset and the `glm` function in R, fit an appropriate Poisson regression model for CHD mortality, adjusting for age, sex and region (eastern/western Finland). Use the first calendar year (1982) as the reference category,

and present the estimated calendar time trend (log-rate ratios) and corresponding 95% confidence intervals graphically. Has the CHD mortality decreased over time? (Note: the < 35 agegroup does not have any CHD deaths, so you may have to modify the data slightly before fitting the model.)

(b) Fit also a similar model for non-CHD mortality (that is, mortality due to causes other that CHD) and comment on whether you can observe a similar trend there.

(c) Instead of log-rate ratios, present the CHD mortality trend over calendar time in terms of the estimated baseline CHD mortality rates, and the corresponding 95% confidence intervals. (For this, you may need to modify the fitted model.)