DS-GA 3001.009 Applied Statistics: Homework #4 Solutions

Due on Thursday, October 19, 2023

Please hand in your homework via Gradescope (entry code: RKXJN2) before 11:59 PM.

1. In class we talked about how to estimate β in the Cox model. This problem investigates the estimation of the baseline survival function $S(t)$ (i.e. the survival function for an individual with $x = 0$).

   (a) Based on the lecture note, explain why the following is a reasonable estimator:

   $$\widehat{S}(t) = \exp\left(- \sum_{i:t_i \le t} \frac{\mathbb{1}(\Delta_i = 1)}{\sum_{k \in R_i} \exp(x_k^\top \widehat{\beta})} \right).$$

   Here R_i is the risk set at time t_i, and $\widehat{\beta}$ is the estimate of β from the Cox model.

   (b) If there is no feature (i.e. $\beta = \widehat{\beta} = 0$), comment on the similarities and differences between the above estimator and the Kaplan-Meier estimator for $S(t)$.

   Solution:

   (a) Let $h(t)$ be the baseline hazard with $H(t) := \int_0^t h(s)ds$, then $S(t) = \exp(-H(t))$ holds in the continuous-time Cox model. Using empirical likelihood (#1 on page 8 of the note), we may assume that $h(\cdot)$ is discrete and $H(t) = \sum_{i:t_i \le t} h(t_i)$. Using the first-order optimality condition for $h(t_i)$ (#2 on page 8 of the note), we have

   $$h(t_i) = \frac{\mathbb{1}(\Delta_i = 1)}{\sum_{k \in R_i} \exp(x_k^\top \beta)}.$$

   Consequently, given the estimate $\widehat{\beta}$ for β, the plug-in approach gives

   $$\widehat{S}(t) = \exp\left(- \sum_{i:t_i \le t} \widehat{h}(t_i) \right) = \exp\left(- \sum_{i:t_i \le t} \frac{\mathbb{1}(\Delta_i = 1)}{\sum_{k \in R_i} \exp(x_k^\top \widehat{\beta})} \right).$$

   (b) When $\beta = 0$, we have

   $$\widehat{S}(t) = \exp\left(- \sum_{i:t_i \le t} \frac{\mathbb{1}(\Delta_i = 1)}{|R_i|} \right) = \exp\left(- \sum_{i:t_i \le t} \frac{d_i}{n_i} \right),$$

   where d_i is the number of observed deaths at time t_i, and n_i is the number of individuals known to have survived right before time t_i (here we assume distinct $t_1, \cdots, t_n$, therefore $d_i \in \{0, 1\}$). In comparison, the Kaplan-Meier estimator is

   $$\widehat{S}_{\text{KM}}(t) = \prod_{i:t_i \le t} \left(1 - \frac{d_i}{n_i} \right).$$

   As $e^{-x} \approx 1 - x$, these two estimates are close to each other. The main difference is that the Kaplan-Meier estimator is derived from a discrete-time model, while $\widehat{S}(t)$ is derived from a continuous-time model.

   (Note: the exponent of $\widehat{S}(t)$ here is called the Nelson-Aalen estimator.)

2. A dataset consists of n observations $(x_1, y_1), \cdots, (x_n, y_n)$, with $x_i \in \mathbb{R}^p, y_i \in \mathbb{N}$, following a multinomial model $(y_1, \cdots, y_n) \sim \text{Multi}(N; (p_1, \cdots, p_n))$ with

$$p_i = \frac{\exp(x_i^\top \beta)}{\sum_{j=1}^n \exp(x_j^\top \beta)}.$$

(a) Show that the log-likelihood under this model is given by $\ell_{\text{M}}(\beta) + c$, where

$$\ell_{\text{M}}(\beta) = \sum_{i=1}^n y_i \left(x_i^\top \beta - \log \left(\sum_{j=1}^n \exp(x_j^\top \beta) \right) \right),$$

and $c \in \mathbb{R}$ is independent of β.

(b) The Poissonization trick introduces an additional parameter $\phi \in \mathbb{R}$ and the following log-likelihood

$$\ell_{\text{P}}(\beta, \phi) = \sum_{i=1}^n \left(y_i (x_i^\top \beta + \phi) - e^{x_i^\top \beta + \phi} \right).$$

Show that ℓ_{M} is the profile likelihood of ℓ_{P}, i.e. $\ell_{\text{M}}(\beta) = \max_{\phi \in \mathbb{R}} \ell_{\text{P}}(\beta, \phi) + c'$ for some constant $c' \in \mathbb{R}$ independent of β.

(c) How does the result in (b) justify the use of Poissonization in Lindsey's method? You may assume $\Delta_k \equiv \Delta$ and $h(z_k) \equiv 1$ in your discussion.

Solution:

(a) Using the multinomial pmf, the log-likelihood is

$$\log \left(\frac{N!}{\prod_{i=1}^n y_i!} \right) + \sum_{i=1}^n \log(p_i^{y_i}) = \ell_{\text{M}}(\beta) + \log \left(\frac{N!}{\prod_{i=1}^n y_i!} \right).$$

(b) Since

$$\frac{\partial \ell_{\text{P}}}{\partial \phi} = n - \sum_{i=1}^n e^{x_i^\top \beta + \phi},$$

the first-order optimality condition gives that

$$\phi = \log n - \log \left(\sum_{i=1}^n e^{x_i^\top \beta} \right).$$

Plugging this expression back to $\ell_{\text{P}}(\beta, \phi)$ gives that

$$\begin{aligned}\max_{\phi \in \mathbb{R}} \ell_{\text{P}}(\beta, \phi) &= \sum_{i=1}^n y_i \left(x_i^\top \beta - \log \left(\sum_{i=1}^n e^{x_i^\top \beta} \right) \right) - e^{\phi} \sum_{i=1}^n e^{x_i^\top \beta} \\ &= \ell_{\text{M}}(\beta) - n.\end{aligned}$$

(c) In Lindsey's method, the original parameter estimation problem is a multinomial regression with log-likelihood ℓ_{M}, while Lindsey's method proposes to solve a Poissonized problem with log-likelihood ℓ_{P} instead; the Poissonized problem becomes a Poisson GLM and has an additional intercept parameter. The result (b) shows that with the help of additional parameter ϕ, maximizing these two likelihoods gives the same parameter β, and therefore explains why Lindsey's method works.

3. Coding: we will explore an AIDS dataset and understand the effects of different treatments on the survival curves for different patients. Based on the inline instructions, fill in the missing codes in `https://tinyurl.com/4bdcyy7c`. Be sure to submit a pdf with your codes, outputs, and colab link.

   Solution: see `https://tinyurl.com/33ndabps`.