DS-GA 3001.009 Applied Statistics: Homework #4

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 $S(t)$.

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 $\ell_{\mathrm{M}}(\beta) + c$, where

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

   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

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

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

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.