Fast Bootstrapping Cox proportional-hazards Model

Shijie Wang

1 Real data analysis

In clinical research, Cox's proportional hazards model is widely used as a semi-parametric regression model to explore the effect of several different risk factors on survival time. For each subject i, the observed response Y_i has the form of min (T_i, C_i) where T_i is the observed response for ith subject and C_i is the censoring time. Let δ_i be censoring indicator for ith subject, such that $\delta = 0$ if the event occurs $(T_i \leq C_i)$ and $\delta = 1$ if the subject is censored $(T_i > C_i)$. The hazard function of Cox model has the form as

$$\lambda(t \mid \mathbf{X}) = \lambda_0(t) \exp(\mathbf{X}^T \theta) \tag{1}$$

where $\lambda_0(t)$ is the unknown baseline hazard rate function. Suppose each subject is independent and censoring time is independent to survival time with no ties in data , θ can be estimated by log partial likelihood, shown below

$$l(\beta) = \sum_{i=1}^{\mathcal{K}} X_i^{\mathrm{T}} \beta - \sum_{i=1}^{\mathcal{K}} \log\big[\sum_{l \in \mathcal{R}(t_i)} e^{X_l^{\mathrm{T}} \beta} \big] \tag{2}$$

where $\mathcal{K}$ is set of index where an event occurs, that is $\delta = 0$, and $\mathcal{R}(t_i)$ is risk set which is defined as $\{X_l : t_l > t_i \text{ for } i = 1, ..., n\}$. In order to implement GBS to fast bootstrap θ in Cox model, the corresponding loss function of GBS is given below,

$$\widehat{G} = \underset{G}{\operatorname{argmin}} \; \mathbb{E}_w \left[\sum_{i=1}^{\mathcal{K}} w_i \cdot \left[X_i^{\mathrm{T}} \cdot G(\mathbf{w}) - \log(\sum_{l \in R(t_i)} e^{X_l^{\mathrm{T}} \cdot G(\mathbf{w})} \right] \right] \tag{3}$$

To examine performance, we implement GBS on a popular lung cancer dataset from the North Central Cancer Treatment Group (NCCTG) (Rubio and Hong, 2016; Rubio and Yu, 2017). NCCTG lung cancer data includes 228 patient observations information from different institutions with their recorded survival time and censor status. NCCTG collected two forms of information from questionnaire filled out by patients and patients'physicians respectively, in purpose of studying independence of descriptive information between patients and their physicians (Loprinzi et al., 1994).

Apart from survival time and censor status, we select other 5 features to our Cox's model, including age in years (age), patient gender (sex), Eastern Cooperative Oncology Group performance score ($ph.ECGO$), Karnofsky performance score rated by physician and patient respectively ($ph.karno$) and Weight loss in last six months ($wt.loss$). For simplicity, we remove 15 rows of records with two or more missing observations in the stage of data cleaning.

For the remaining dataset (size $n = 213$), we consider a multivariate cox regression model with 7 features described above and obtain GBS distribution with bootstrap sample size 10,000. Results are summarized in figure 13, where MLE bootstrap distribution of θ is obtained by nonparametric bootstrapping. The GBS bootstrap distribution of θ is nearly indistinguishable to conventional MLE bootstrap distribution under Cox's model.

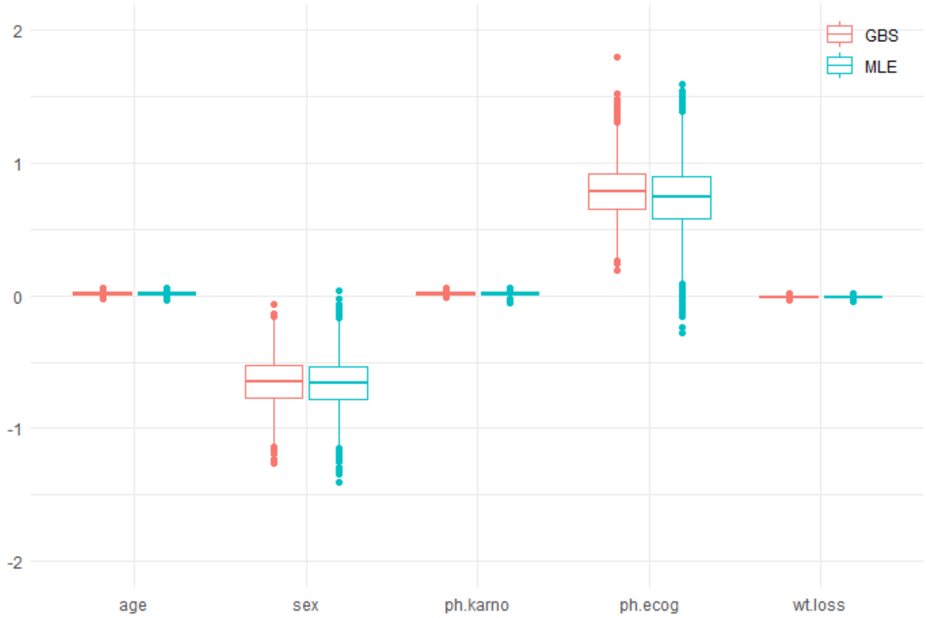

Figure 1: Bootstrap distribution comparison between GBS and conventional nonparametric bootstrap