Survival analysis

Badr MOUFAD

May 30, 2023

Contents

1 Cox datafit

Let's $\mathbf{X} \in \mathbb{R}^{n\times p}$ be a matrix of p predictors and n samples $x_i \in \mathbb{R}^p$, $y \in \mathbb{R}^n$ a vector recording the time of events occurrences, and $s \in \{0,1\}^n$ a binary vector where 1 means *event occurred*, and finally $\beta \in \mathbb{R}^p$ the vector of coefficient to be estimated.

1.1 Breslow

Proposition 1. *[Lin, 2007, Section 2] The expression of the negative log-likelihood according to Breslow estimate reads*

$$l(\beta) = \sum_{i=1}^{n} -s_i \langle x_i, \beta\rangle + s_i \log(\textstyle\sum_{y_j \geq y_i} e^{\langle x_j, \beta\rangle}) \ \ . \tag{1}$$

To get a more compact expression, we introduce the matrix $\mathbf{B} \in \mathbb{R}^{n\times n}$ defined as

$$\mathbf{B}_{i,j} = \mathbb{1}_{y_j \geq y_i} = \begin{cases} 1, & \text{if } y_j \geq y_i, \\ 0, & \text{otherwise} \end{cases} \ \ , \tag{2}$$

and we let b_i be its i-th row.

Proposition 2. *The expression in Equation (1) is equivalent to*

$$l(\beta) = -\langle s, \mathbf{X}\beta\rangle + \langle s, \log(\mathbf{B}e^{\mathbf{X}\beta})\rangle \ \ . \tag{3}$$

Proof. We can observe that the sum can be split into two parts. The first one,

$$\sum_{i=1}^{n} -s_i \langle x_i, \beta\rangle = -\langle s, \mathbf{X}\beta\rangle \ \ .$$

And the second part,

$$\begin{aligned} \sum_{i=1}^{n} s_i \log(\textstyle\sum_{y_j \geq y_i} e^{\langle x_j, \beta\rangle}) &= \sum_{i=1}^{n} s_i \log(\textstyle\sum_{j=1}^{n} \mathbb{1}_{y_j \geq y_i} e^{\langle x_j, \beta\rangle}) \\ &= \sum_{i=1}^{n} s_i \log(\textstyle\sum_{j=1}^{n} b_{ij} e^{\langle x_j, \beta\rangle}) \\ &= \sum_{i=1}^{n} s_i \log\langle b_i, e^{\mathbf{X}\beta}\rangle \\ &= \langle s, \log(\mathbf{B}e^{\mathbf{X}\beta})\rangle \ \ . \end{aligned}$$

Combining the two expressions we get the desired expression. □

The latter defines the Cox datafit. We note that it only depends on $\mathbf{X}\beta$. Indeed, by considering $F : \mathbb{R}^n \to \mathbb{R}$ such that

$$F(u) = -\langle s, u\rangle + \langle s, \log(\mathbf{B}e^{u})\rangle \ \ , \tag{4}$$

it follows that $l(\beta) = F(\mathbf{X}\beta)$. Therefore, from now on, we focus on F to derive the gradient and Hessian of the datafit.

Proposition 3. *For some u in $\mathbb{R}^n$, the gradient of F reads*

$$\nabla F(u) = -s + [\operatorname{diag}(e^u)\mathbf{B}^\top]\frac{s}{\mathbf{B}e^u} \ ,$$

where the fraction is performed element-wise.

Proof. Deferred in the appendix. □

Proposition 4. *For some u in $\mathbb{R}^n$, the Hessian of F is*

$$\nabla^2 F(u) = \operatorname{diag}(e^u \odot \mathbf{B}^\top \frac{s}{\mathbf{B}e^u}) - \operatorname{diag}(e^u)\mathbf{B}^\top \operatorname{diag}(\frac{s}{(\mathbf{B}e^u)^2})\mathbf{B}\operatorname{diag}(e^u) \ , \tag{5}$$

where the square and fraction operations are performed element-wise.

Proof. Deferred in the appendix. □

The Hessian, as it is, is costly to evaluate because of the right hand-side term. Indeed, the latter involves, in particular, a $\mathcal{O}(n^3)$ operation. We overcome this limitation thanks to the proposition below.

Proposition 5. *For some u in $\mathbb{R}^n$, the Hessian in Equation (5) can be overestimated as follows*

$$\nabla^2 F(u) \preccurlyeq \operatorname{diag}(e^u \odot \mathbf{B}^\top \frac{s}{\mathbf{B}e^u}) \ .$$

Proof. We have to show that $\operatorname{diag}(e^u \odot \mathbf{B}^\top \frac{s}{\mathbf{B}e^u}) - \nabla^2 F(u) := \mathbf{\Phi}$ is positive semi-definite.

Let u be in $\mathbb{R}^n$, we have

$$\begin{aligned} \langle \mathbf{\Phi} u, u \rangle &= \left\langle \operatorname{diag}(e^u)\mathbf{B}^\top \operatorname{diag}(\frac{s}{(\mathbf{B}e^u)^2})\mathbf{B}\operatorname{diag}(e^u)u, u \right\rangle \\ &= \left\| \operatorname{diag}(\frac{\sqrt{s}}{\mathbf{B}e^u})\mathbf{B}\operatorname{diag}(e^u)u \right\|^2 \geq 0 \ , \end{aligned}$$

which enables us to conclude. □

1.2 Efron

Efron estimate refines Breslow by handling tied observations, *observations with identical occurrences' time*. Let's define $H_k = \{i \mid s_i = 1 \ ; \ y_i = y_k\}$, the set of uncensored observations with the same occurrence time y_k, and denote $y_{i_1}, \ldots, y_{i_m}$ the unique times, assumed to be in total equal to m.

Proposition 6. *[Efron, 1977, Section 6, equation (6.7)] The minus log-likelihood according to Efron estimate is*

$$l(\beta) = \sum_{l=1}^{m} \left(\sum_{i \in H_{i_l}} -\langle x_i, \beta \rangle + \log\Big(\textstyle\sum_{y_j \geq y_{i_l}} e^{\langle x_j, \beta \rangle} - \frac{\#(i)-1}{|H_{i_l}|} \sum_{j \in H_{i_l}} e^{\langle x_j, \beta \rangle} \Big) \right) \ , \tag{6}$$

where $|H_{i_l}|$ stands for the cardinal of H_{i_l}, and $\#(i)$ the index of observation i in H_{i_l}.

Ideally, we would like to rewrite this expression like Equation (3) to leverage the established results about the gradient and Hessian. What distinguishes both expressions is the presence of a double sum and second term within the log.

Proposition 7. *The expression in Equation (6) is equivalent to*

$$l(\beta) = -\langle s, \mathbf{X}\beta\rangle + \langle s, \log(\mathbf{B}e^{\mathbf{X}\beta} - \mathbf{A}e^{\mathbf{X}\beta})\rangle \ , \tag{7}$$

where $\mathbf{A}$ *is linear operator defined by Algorithm 1.*

Proof. First, we can observe that $\cup_{l=1}^{m} H_{i_l} = \{i \mid s_i = 1\}$, which enables us to write the double sum as a single one. Therefore,

$$\begin{aligned}\sum_{l=1}^{m}\sum_{i\in H_{i_l}} -\langle x_i, \beta\rangle = \sum_{i:s_i=1} -\langle x_i, \beta\rangle &= \sum_{i=1}^{n} -s_i\langle x_i, \beta\rangle \\ &= -\langle s, \mathbf{X}\beta\rangle \ .\end{aligned}$$

On the other hand, we have

$$\begin{aligned}-\frac{\#(i)-1}{|H_{i_l}|}\sum_{j\in H_{i_l}} e^{\langle x_j,\beta\rangle} &= \sum_{j=1}^{n} -\frac{\#(i)-1}{|H_{i_l}|}\ \mathbb{1}_{j\in H_{i_l}}\ e^{\langle x_j,\beta\rangle} \\ &= \sum_{j=1}^{n} a_{i,j}e^{\langle x_j,\beta\rangle} \\ &= \langle a_i, e^{\mathbf{X}\beta}\rangle \ ,\end{aligned}$$

where a_i is a vector in $\mathbb{R}^n$ chosen accordingly to preform the linear operation.

Defining the matrix $\mathbf{A}$ with rows $(a_i)_{i\in[n]}$, and combining that with the first result enable us to conclude. □

Since the expression of the gradient and Hessian involve the adjoint of $\mathbf{A}$, we present also Algorithm 2 to evaluate $\mathbf{A}^\top v$, for some v in $\mathbb{R}^n$.

We notice that the complexity of both algorithms is $\mathcal{O}(n)$ despite intervening a matrix multiplication. This is due to the special structure of $\mathbf{A}$ which in the case of sorted observations has a block diagonal structure with each block having equal columns.

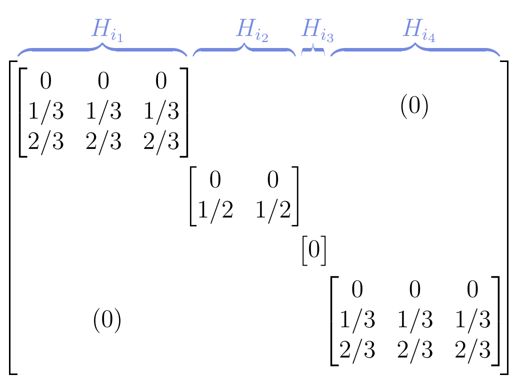

Figure 1: Strucutre of $\mathbf{A}$ in the case of sorted observations with group sizes of identical occurrences times being $3, 2, 1, 3$ respectively.

Algorithm 1 Evaluate $\mathbf{A}v$

input: $v \in \mathbb{R}^n$
init **:** $o \in \mathbb{R}^n$
for $l = 1, \cdots, m$ **do**
  $\bar{v}_{H_{i_l}} \leftarrow \text{sum}(v_{H_{i_l}})$
  $o_{H_{i_l}} \leftarrow \bar{v}_{H_{i_l}} \times [0, \frac{1}{|H_{i_l}|}, \ldots, \frac{|H_{i_l}|-1}{|H_{i_l}|}]$
return o

Algorithm 2 Evaluate $\mathbf{A}^\top v$

input: $v \in \mathbb{R}^n$
init **:** $o \in \mathbb{R}^n$
for $l = 1, \cdots, m$ **do**
  $w_{H_{i_l}} \leftarrow \langle v_{H_{i_l}}, [0, \frac{1}{|H_{i_l}|}, \ldots, \frac{|H_{i_l}|-1}{|H_{i_l}|}] \rangle$
  $o_{H_{i_l}} \leftarrow w_{H_{i_l}} \times [1, \ldots, 1]$
return o

References

DY Lin. On the breslow estimator. *Lifetime data analysis*, 13:471–480, 2007.

Bradley Efron. The efficiency of cox's likelihood function for censored data. *Journal of the American statistical Association*, 72(359):557–565, 1977.