**Cox model with distributed data using the WebDISCO algorithm**

Estimates and confidence intervals for the parameters of the Cox model are obtained by making use of the Newton-Raphson algorithm using local gradients and Hessian matrices. This allows for the recreation of the estimates and confidence intervals from the centralised setting. For the procedure, all nodes must have the same predictors.

The initial step consists of each node sharing all their unique event times to the coordination center. The global server then identifies and consolidates all unique event times across all sites. Each site then initializes parameters that remain constant throughout the process. Next, the coordination center initializes the first beta with an estimate of their choice.

The iterative process starts with each site computing data aggregates. These aggregates are then used by the central server to calculate gradients and Hessian matrices, which in turn are used to update the parameter estimates. This iterative process continues until convergence is achieved (which yields the centralised estimate).

In the following procedure, is the site number, is the current iteration number ( and represents the number of sites . is the number of observations in a site, is the number of predictors, represents the event time number, and represents the total number of distinct event time . and are the indices of the element in the parameter vector , is the variable (predictors) for an individual subject . is the optional local vector of weights. If no weights are provided, a vector of 1s will be used instead, as this represents uniform weights across observations.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Example.**  Suppose the following dataset at node , with 5 observations () and 2 predictors ():   |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | | **Id** | **time** | **status** | **age** | **sex** |  | **weights** | | 1 | 3 | 1 | 42 | 1 |  | 2 | | 2 | 6 | 0 | 38 | 1 |  | 1 | | 3 | 11 | 1 | 37 | 2 |  | 3 | | 4 | 11 | 1 | 51 | 1 |  | 2 | | 5 | 14 | 1 | 36 | 2 |  | 5 |   where time represents the time of event occurrence or censoring, status indicates whether the event has occurred (1) or the data is censored (0) and age and sex are the predictors. The last column is the optional weight vector, which is provided in this example. |

**Data node (initial phase)**

1. Each node identifies unique event times and saves this data to a CSV file (Times\_k\_output.csv). Each site also computes their local Cox models to obtain local betas and variance-covariance matrices, saving them to a CSV file (Beta\_local\_k.csv) along with the number of observations (N\_node\_k.csv). For verification purposes, all predictor variable names are also saved in a CSV file (Predictor\_names\_k.csv). All files are then sent to the coordinating node.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Example (continued).**  Estimates for the Cox proportional hazard model can be found using functions such as *coxph* in R, and we obtain:  The following data is **shared** to the coordinating node:   |  | | --- | | **time** | |  | |  | |  |   , , and . |

**Coordinating node (initial phase)**

1. The coordinating node gathers all unique times from every site, generates a list of unique times across all sites, then sends it back to the data nodes (Global\_times\_output.csv). The global server also computes the inverse variance weighted initial estimator, , and saves it to a CSV file.[[1]](#footnote-2)

**Data node (first iteration)**

1. Each node initialises the following parameters: (Rikk.csv), a list containing the IDs of the at-risk subjects for all times , (Rik\_compk.csv), a list containing the IDs of subjects that died at or by time and , a list containing the IDs of subjects that had an event at time . Those files will be used by the node throughout the analysis.   
     
   Each node also computes the following: (sumWZrk.csv), the weighted sum for a component of the predictor for subjects with an event at time , (Wprimek.csv) and the list containing the total weight of events at time for node . All of these CSV files are sent to the coordinating node.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Example (continued).**  The following data is computed but **not shared** to the coordinating node:   |  |  |  |  | | --- | --- | --- | --- | | *i* |  |  |  | |  |  |  |  | |  |  |  |  | |  |  |  |  |   The following data is **shared** to the coordinating node:   |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | | |  |  | | --- | --- | | *i* |  | |  |  | |  |  | |  |  | | |  |  |  | | --- | --- | --- | | *i* |  | | | ***r = 1*** | ***r = 2*** | |  |  |  | |  |  |  | |  |  |  | | |

**Data node (following iterations)**

**4a.** Then, each data node computes aggregated statistics used for the gradient and Hessian:

* , a matrix (sumWExpk\_output\_t.csv)
* , a matrix (sumWZqExpk\_output\_t.csv)
* , a matrix (sumWZqZrExpk\_output\_t.csv)

These quantities are sent to the coordinating center using CSV files.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Example (continued).**  Suppose the coordinating node shared .  For the first event time :        The following quantities are **shared** to the coordinating node:   |  |  | | --- | --- | | *i* |  | |  | 0.031 | |  | 0.0086 | |  | 0.0011 |  |  |  |  | | --- | --- | --- | | *i* |  | | | ***q = 1*** | ***q = 2*** | |  | 1.295 | 0.0331 | |  | 0.406 | 0.0108 | |  | 0.0401 | 0.0022 |  |  |  |  |  | | --- | --- | --- | --- | | *i* | |  | | | ***q = 1*** | ***q = 2*** | |  | ***r = 1*** | 55.02 | 1.374 | | ***r = 2*** | 1.374 | 0.037 |  |  |  |  |  | | --- | --- | --- | --- | |  | ***r = 1*** | 19.56 | 0.485 | | ***r = 2*** | 0.485 | 0.015 |  |  |  |  |  | | --- | --- | --- | --- | |  | ***r = 1*** | 1.443 | 0.080 | | ***r = 2*** | 0.080 | 0.0045 | |

**4b.** When using *case weights*, Therneau & Grambsch[[2]](#footnote-3) suggest using a robust variance estimation, based on score residuals[[3]](#footnote-4). In order to do so, more intermediate statistics must be shared between nodes and coordinating node.

All quantities in section 4**b** are required to compute the score residuals in a distributed way.

**(Iterations t>1)**

Each data node computes aggregated statistics used for the robust variance estimation:

(zbarri\_inverseWExp\_k\_output\_(t-1).csv),

and

(inverseWExp\_k\_output\_(t-1).csv).

The CSV files are then shared to the coordinating node.

**(Iterations t>2)**Each node should now be able to compute the score residuals of their subjects. Using this, each node will compute the following aggregated statistic used for the robust variance estimation.

First, each node will compute their subjects’ *Schoenfeld residuals*:

where is the event indicator for subject .

Afterwards, each node will compute their subjects’ *score residuals*:

Finally, each node will compute their contribution to the estimation of the jackknife estimate of variance , where:

Note that only the diagonal of the matrix will be shared to the coordinating node through a CSV (DDk\_output\_(t-2).csv) as non diagonal entries are left unused.

The CSV files are then shared to the coordinating node.

**Coordinating node (first iteration only)**

1. The global server aggregates (sumWZrGlobal.csv), which is the sum of all predictors for subjects that had an event across all sites, and calculates (WprimeGlobal.csv), the list containing the total weight of events at time .

**Coordinating node (following iterations)**

**6a.** The coordinating center computes the global gradient and Hessian:

It then computes the Newton-Raphson iteration and sends the new estimate to the data nodes.

Steps **4a.** and **6a.** are repeated manually until convergence.

**6b.** When using *case weights*, Therneau & Grambsch suggest using a robust variance estimation, based on score residuals. In order to do so, more intermediate statistics must be shared between nodes and coordinating node.

All quantities in section **6b** are required to compute the score residuals in a distributed way.

**(Iterations t>1)**

The coordinating center computes and stores in a CSV file both:

(sumWExpGlobal\_output\_(t-1).csv),

and

(zbarri\_(t-1).csv).

The CSV files are then shared to the local nodes.

**(Iterations t>2).**

The coordinating center computes both

(zbarri\_inverseWExp\_Global\_output\_(t-2).csv),

and

(inverseWExp\_Global\_output\_(t-2).csv).

The CSV files are then shared to the local nodes.

Steps **4b.** and **6b.** are repeated manually until convergence.

**7a.** For all iterations except the first, the coordinating node computes the confidence intervals. The (estimated) covariance matrix of the coefficients is . Said matrix will be saved as “Fisher\_t.csv” and may be shared with local nodes in order to get robust variance estimates.

Lower and upper bounds for the confidence intervals of the model parameters are also calculated at the coordinating node:

The outputs of the procedure are the coefficients, the upper and lower bounds of the confidence intervals for the exponential of the model parameters, the standard error and the p values, for the previous iteration.

*Final results will be located in the Results\_t.csv file.*

**7b.** When using *case weights*, Therneau & Grambsch suggest using a robust variance estimation, based on score residuals. In order to do so, more intermediate statistics must be shared between nodes and coordinating node.

**(Iterations t>3)**

The coordinating node has now all the needed information to compute the robust standard error of . These standard errors are the squared root of the diagonal of the variance matrix . This allows the coordinating node to produce a new results file (RobustResults\_(t-3).csv) which is an updated version of Results\_(t-3).csv.

*Final results will be located in the RobustResults\_t.csv file.*

1. Duan, R., Luo, C., Schuemie, M. J., Tong, J., Liang, C. J., Chang, H. H., ... & Chen, Y. (2020). Learning from local to global: An efficient distributed algorithm for modeling time-to-event data. *Journal of the American Medical Informatics Association*, *27*(7), 1028-1036. [↑](#footnote-ref-2)
2. T.M. Therneau and P. M. Grambsch. *Modeling Survival Data: Extending the Cox Model.* Statistics for Biology and Health. Springer New York, 2013. [↑](#footnote-ref-3)
3. Collett, D. (2023). Modelling Survival Data in Medical Research (4th ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9781003282525 [↑](#footnote-ref-4)