Confidentiality for Cox model

The WebDISCO method for estimating a survival model does not ensure data confidentiality because predictors can be retrieved by the central server. This occurs due to sending data at every event time. When only one event happens during a given period, only the predictors of that individual vary, making it possible to identify them.

To address this issue, it is important to ensure that there is never a single event occurring at a single event time. Therefore, data can be divided into intervals where many patients are grouped. Various methods are available for this purpose.

Averaging

The data is ordered by time, and values are grouped and given a new time, which is the average of all values of time.

Example.

For the example, we assume we want to group data into groups of 2.

time	p_1	p_2
2	43	0
4	24	0
5	41	1
6	37	1
9	53	0
11	33	1
12	39	1
17	45	0

Before

time	p_1	p_2
3	43	0
3	24	0
5.5	41	1
5.5	37	1
10	53	0
10	33	1
14.5	39	1
14.5	45	0

After

Where p_1 and p_2 are predictors.

Uniform Intervals (with cutoff)

The period of the study is split into uniform intervals that contain at least x subjects per interval. The last few values are excluded when choosing the interval size, as they can be spread far apart.

Example

A percentage of values to exclude is chosen. In this example, we choose 15%, which results in the exclusion of the last value, 17.

To ensure there are at least 2 values in every interval, the intervals must be larger than the largest difference between any two consecutive time values. In this case, the largest difference is calculated as $9 - 6 + 1 = 4$. Therefore, the intervals should be of size 4.

time	p_1	p_2
2	43	0
4	24	0
5	41	1
6	37	1
9	53	0
11	33	1
12	39	1
17	45	0

Before

time	p_1	p_2
3.5	43	0
3.5	24	0
3.5	41	1
7.5	37	1
7.5	53	0
11.5	33	1
11.5	39	1
11.5	45	0

After

Intervals: 2 – 5, 6 – 9, 10 – 13.
The last value, 17, will go in the last interval.

Non-uniform Intervals

The period of the study is split into non-uniform intervals. These intervals are the smallest possible size that contains x subjects.

Example

Each interval should contain the minimum number of values possible. Ideally, each interval will contain exactly 2 values.

time	p_1	p_2
2	43	0
4	24	0
5	41	1
6	37	1
9	53	0
11	33	1
12	39	1
17	45	0

Before

time	p_1	p_2
3	43	0
3	24	0
5.5	41	1
5.5	37	1
10	53	0
10	33	1
14.5	39	1
14.5	45	0

After

Intervals: 2 – 4, 5 – 6, 9 – 11, 12 – 17.
However, all the sites must agree on these intervals.