compSurv: A Complete Survival Analysis Tool

Selcuk Korkmaz

November 27, 2016

Contents

Abstract

Survival analysis is described as collection of statistical methods for which the response variable of interest is time until an event occurs. In this context, the time can be days, week, months and years from the beginning of follow-up of an individual until an event occurs, or the age of an individual when the event occurs. Moreover, the event can be death, disease, remission, recovery or any experience of interest that may occur to an individual. A more detailed information can be found in Kleinbaum and Marubini and Valsecchi.

Here we developed an easy-to-use, up-to-date, comprehensive and interactive web-based tool for survival analysis. This tool includes analysis procedures for life table, Kaplan-Meier and Cox regression. Each procedure includes following features:

Life table: descriptive statistics, life table, median life time, hazard ratios and comparison tests including Log-rank, Gehan-Breslow, Tarone-Ware, Peto-Peto, Modified Peto-Peto, Flemington-Harrington.

Kaplan-Meier: descriptive statistics, survival table, mean and median life time, hazard ratios, comparison tests including Log-rank, Gehan-Breslow, Tarone-Ware, Peto-Peto, Modified Peto-Peto, Flemington-Harrington, and interactive plots such as Kaplan-Meier curves and hazard plots.

Cox regression: coefficient estimates, hazard ratios, goodness of fit test, analysis of deviance, save predictions, save residuals, save Martingale residuals, save Schoenfeld residuals, save dfBetas, proportional hazard assumption test, and interactive plots including Schoenfeld residual plot and Log-Minus-Log plot.

Regularized Cox regression: variable selection and coefficient estimations using ridge, elastic net and lasso penalties.

Random survival forests: individual survival and cumulative hazard predictions using random survival forests, and interactive plots including, survival (with OOB), hazard (with OOB), error rate vs number of tree and cox regression vs random survival forest model.

1.Data upload

This tool requires a dataset in `*.txt` format, which is seperated by `comma`, `semicolon`, `space` or `tab` delimiter. First row of dataset must include header. When the appropriate file is uploaded, the dataset will be appear immediately on the main page of the tool. Alternatively users can upload one of the example datasets provided within the tool for testing and understanding the operating logic of the tool.

2. Analysis Methods

2.1. Life Table

A life table can be constructed for a cohort group using the following steps:

1. Select the analysis method as `Life Table` from `Analysis` tab.
2. Select suitable variables for the analysis, such as `survival time`, `status variable`, `category value for status variable` and `factor variable`, if exists.
3. Define an appropriate time interval from beginning to end of study by a specific step.
4. In advanced options, one can change `confidence interval type` as log, log-log or plain, `variance estimation method` as Greenwood or Tsiatis and `reference category` as first or last.
5. Click `Run` button to run the analysis.

Desired outputs can be selected by clicking Outputs checkbox. Available outputs are;

---

a. Case summary

Copy CSV PDF Print Search:

	n	n of event	% of event	n of censor	% of censor
1	48	29	60.417	19	39.583
2	38	18	47.368	20	52.632

Showing 1 to 2 of 2 entries Previous 1 Next

---

b. Life table

Copy CSV PDF Print Search:

	interval time	n of subjects	n of withdraws	n of subjects at risk	n of events	prob. of terminating during the time interval	prob. of surviving in the time interval	cum. prob. of surviving at the begining of the time interval
1.1	0_<10	48	4	46	19	0.413	0.587	1
1.2	10_<20	25	3	23.5	6	0.255	0.745	0.587
1.3	20_<30	16	5	13.5	3	0.222	0.778	0.437
1.4	30_<40	8	4	6	1	0.167	0.833	0.340
1.5	40_<50	3	2	2	0	0.000	1.000	0.283
1.6	50_<60	1	1	0.5	0	0.000	1.000	0.283
2.1	0_<10	38	3	36.5	12	0.329	0.671	1
2.2	10_<20	23	3	21.5	2	0.093	0.907	0.671
2.3	20_<30	18	4	16	3	0.188	0.812	0.609
2.4	30_<40	11	1	10.5	1	0.095	0.905	0.495

Showing 1 to 10 of 12 entries Previous 1 2 Next

---

c. Median life time

Copy CSV PDF Print Search:

	Median life time
1	15.802
2	29.531

Showing 1 to 2 of 2 entries Previous 1

c. Hazard ratio

Copy CSV PDF Print Search:

	Interval time	HR	S.E.	Lower limit	Upper limit
1.1	0_<10	0.052	0.012	0.029	0.075
1.2	10_<20	0.029	0.012	0.006	0.052
1.3	20_<30	0.025	0.014	-0.003	0.053
1.4	30_<40	0.018	0.018	-0.017	0.054
1.5	40_<50	0			
1.6	50_<60				
2.1	0_<10	0.039	0.011	0.018	0.061
2.2	10_<20	0.01	0.007	-0.004	0.023
2.3	20_<30	0.021	0.012	-0.003	0.044
2.4	30_<40	0.01	0.01	-0.01	0.03

Showing 1 to 10 of 12 entries Previous 1 2 Next

c. Comparison tests

Copy CSV PDF Print Search:

	Test	Chi square	df	p value
1	Log-rank	1.489	1	0.222
2	Gehan-Breslow	0.744	1	0.388
3	Tarone-Ware	1.092	1	0.296
4	Peto-Peto	0.888	1	0.346
5	Modified Peto-Peto	0.874	1	0.35
6	Flemington-Harnington	2.606	1	0.106

Showing 1 to 6 of 6 entries Previous 1

2.2. Kaplan-Meier

A Kaplan-Meier analysis can be conducted by applying the following steps:

1. Select the analysis method as `Kaplan Meier` from `Analysis` tab.
2. Select suitable variables for the analysis, such as `survival time`, `status variable`, `category value for status variable` and `factor variable`, if exists.
3. In advanced options, one can change confidence interval type, as log, log-log or plain, variance estimation method, as Greenwood or Tsiatis, comparison test type, as Log-rank, Gehan-Breslow, Tarone-Ware,

Peto-Peto, Modifi Peto-Peto or Flemington-Harrington, confidence level and reference category, as first or last.

4. Click `Run` button to run the analysis.

Desired outputs can be selected by clicking Outputs checkbox. Available outputs are;

---

a. Case summary

Copy CSV PDF Print Search:

	n	n of event	% of event	n of censor	% of censor
1	48	29	60.417	19	39.583
2	38	18	47.368	20	52.632

Showing 1 to 2 of 2 entries Previous 1 Next

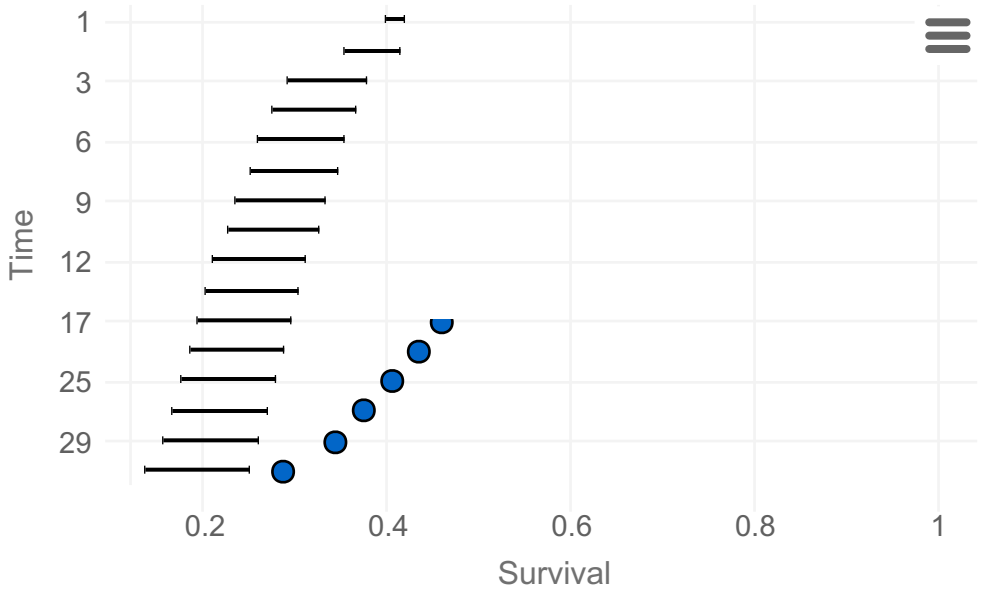
---

b. Survival table

Copy CSV PDF Print Search:

	Time	Number at risk	Number of event	Cumulative probability of surviving	S.E.	Lower limit	Upper limit
1.1	1	47	1	0.979	0.021	0.938	
1.2	2	45	4	0.892	0.046	0.806	0.9
1.3	3	41	7	0.739	0.065	0.623	0.8
1.4	5	33	2	0.695	0.068	0.573	0.8
1.5	6	31	2	0.65	0.071	0.525	0.8
1.6	7	29	1	0.627	0.072	0.502	0.7
1.7	9	27	2	0.581	0.074	0.453	0.7
1.8	10	25	1	0.558	0.074	0.43	0.7
1.9	12	23	2	0.509	0.075	0.381	0.
1.10	16	20	1	0.484	0.076	0.356	0.6

Showing 1 to 10 of 27 entries Previous 1 2 3 Next



c. Survival plot

Copy CSV PDF Print Search:

	n	Mean	S.E.(mean)	Lower (mean)	Upper (mean)	Median	Lower (median)
1	48	25.127	3.7	17.875	32.378	16	9
2	38	32.932	4.352	24.403	41.46	26	17

Showing 1 to 2 of 2 entries Prev

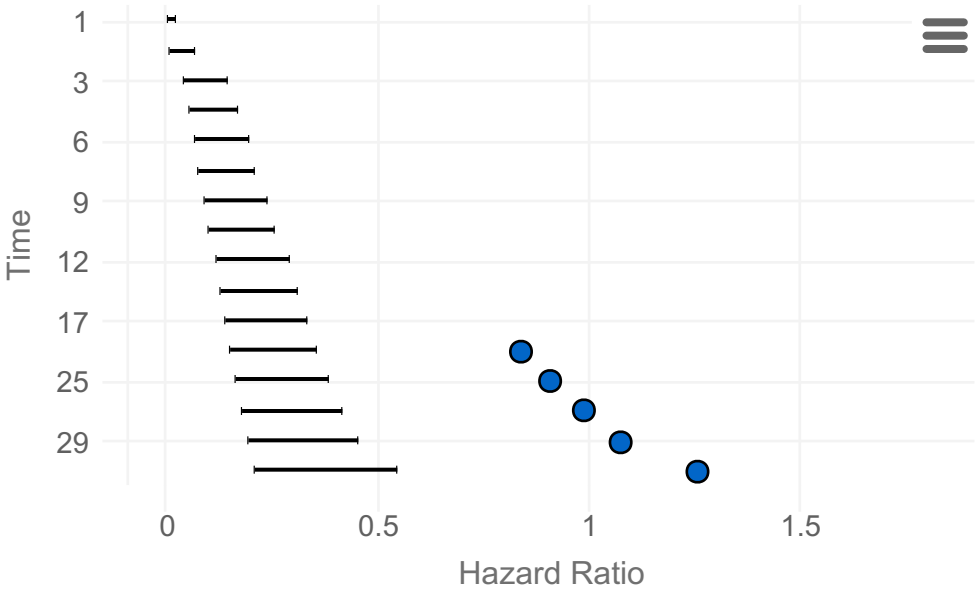
d. Mean and Median life time

Copy CSV PDF Print Search:

	Time	Hazard.Ratio	Lower	Upper
1.1	1	0.021	0	0.064
1.2	2	0.114	0.014	0.216
1.3	3	0.302	0.13	0.473
1.4	5	0.364	0.172	0.557
1.5	6	0.431	0.218	0.644
1.6	7	0.467	0.242	0.689
1.7	9	0.543	0.294	0.792
1.8	10	0.583	0.323	0.844
1.9	12	0.675	0.386	0.965
1.10	16	0.726	0.42	1.033

Showing 1 to 10 of 27 entries Previous 1 2 3 Next

e. Hazard ratio



e. Hazard plot

Copy CSV PDF Print Search:

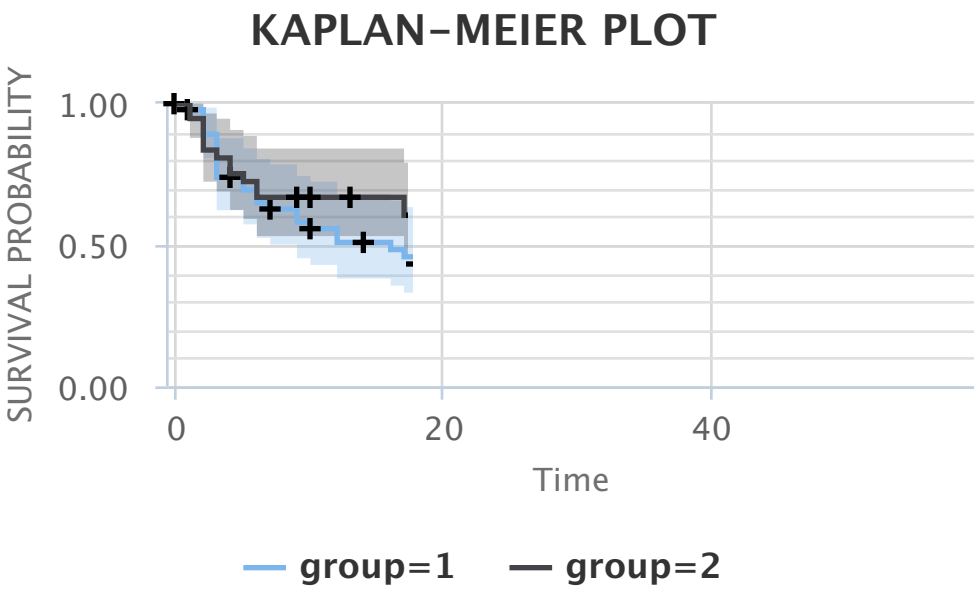
	Test	Chi square	df	p value
1	Log-rank	1.521	1	0.217
2	Gehan-Breslow	0.923	1	0.337
3	Tarone-Ware	0.663	1	0.415
4	Peto-Peto	0.888	1	0.346
5	Modified Peto-Peto	0.874	1	0.35
6	Flemington-Harrington	2.606	1	0.106

Showing 1 to 6 of 6 entries Previous 1

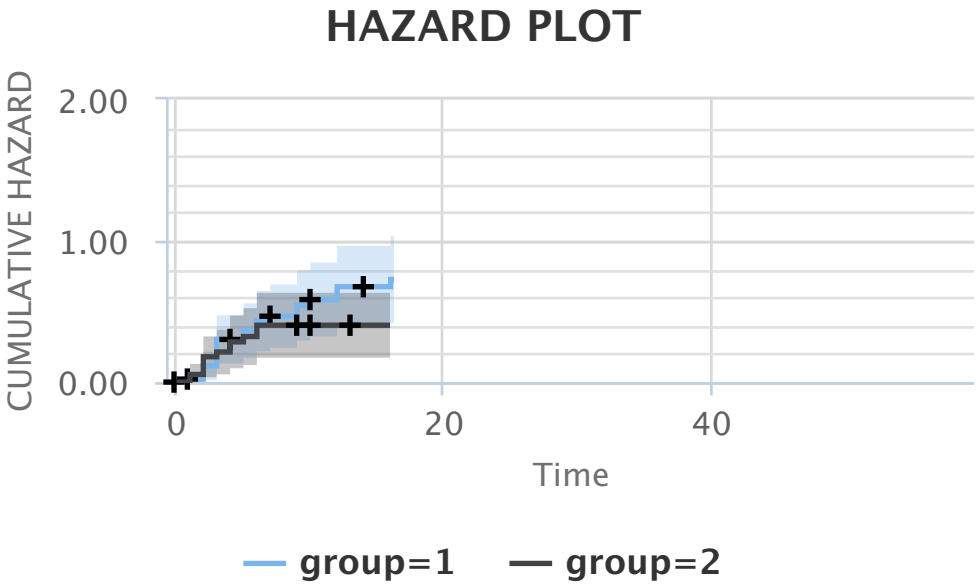
g. Comparison tests

h. Plots

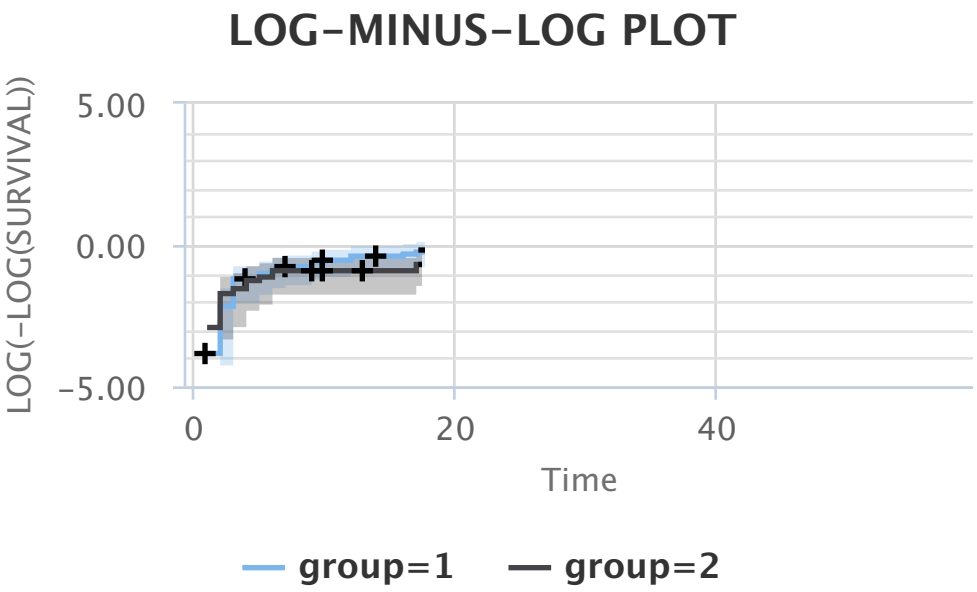
Kaplan-Meier curve



Hazard plot



Log-Minus-Log plot



2.3. Cox Regression

A Cox regression analysis can be conducted by applying the following steps:

1. Select the analysis method as `Cox Regression` from `Analysis` tab.
2. Select suitable variables for the analysis, such as `survival time`, `status variable`, `category value for status variable`, and categorical and continuous predictors for the model.
3. In advanced options, `interaction terms`, `strata terms` and `time dependent covariates` can be added to the model. Moreover, if there are multiple records for observations, users can specify it by clicking `Multiple ID` checkbox. Furthermore, once can choose model selection criteria, as `AIC` or `p-value`, model selection method, as `backward`, `forward` or `stepwise`, reference category, as `first` or `last`, and ties method, as `Efron`, `Breslow` or `exact` and change the `confidence level`.
4. Click `Run` button to run the analysis.

Desired outputs can be selected by clicking Outputs checkbox. Available outputs are coefficient estimates, hazard ratio, goodness of fit tests, analysis of deviance, predictions, residuals, Martingale residuals, Schoenfeld residuals and DfBetas.

a. Coefficient Estimates

Copy CSV PDF Print Search:

	Variable	Coefficient estimate	Standard error	z statistic	
1	group1	0.553	0.448	1.235	0.2
2	age	0.150	0.036	4.107	0.0

Showing 1 to 2 of 2 entries Previous

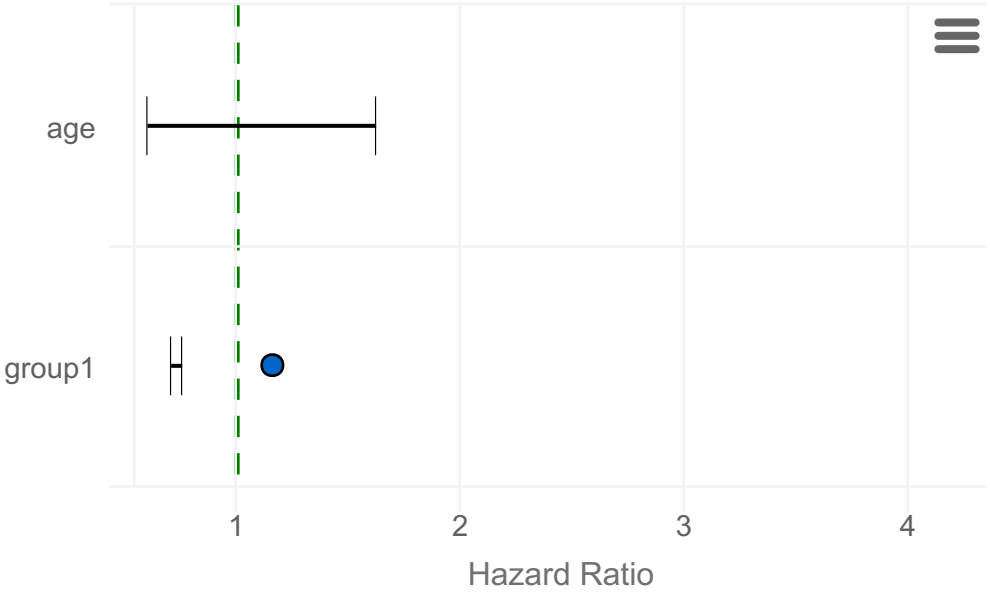
b. Hazard ratio

Copy CSV PDF Print Search:

	Variable	Hazard ratio	Lower limit (95%)	Upper limit (95%)
1	group1	1.738	0.723	4.178
2	age	1.161	1.081	1.247

Showing 1 to 2 of 2 entries Previous 1 Next

c. Hazard plot



d. Goodness of Fit Tests

Copy CSV PDF Print Search:

Test	Statistic	df	p value
Likelihood ratio	23.198	2	0.000
Wald	17.190	2	0.000
Score (logrank)	21.164	2	0.000

Showing 1 to 3 of 3 entries Previous 1 Next

******

e. Analysis of Deviance

Copy CSV PDF Print Search:

Variable	df	Chisquare	p value
group	1	0.000	0.983
age	1	23.197	0.000

Showing 1 to 2 of 2 entries Previous 1 Next

******

f. Predictions

Copy CSV PDF Print Search:

	Predictions
1	0.169
2	0.169
3	0.169
4	0.169
5	0.169
6	0.169
7	0.169
8	0.169
9	0.169
10	0.169

Showing 1 to 10 of 402 entries Previous 1 2 3 4 5 … 41 Next

******

g. Residuals

Copy CSV PDF Print Search:

	Residuals
1	-0.130
2	-0.147
3	-0.161
4	-0.163
5	-0.171
6	-0.186
7	-0.192
8	-0.214
9	-0.216
10	-0.264

Showing 1 to 10 of 402 entries Previous 1 2 3 4 5 … 41 Next

******

h. Martingale Residuals

Copy CSV PDF Print Search:

	Martingale Residuals
1	-0.008
2	-0.011
3	-0.013
4	-0.013
5	-0.015
6	-0.017
7	-0.018
8	-0.023
9	-0.023
10	-0.035

Showing 1 to 10 of 402 entries Previous 1 2 3 4 5 … 41 Next

******

i. Schoenfeld Residuals

Copy CSV PDF Print Search:

	group1	age
1	0.291	6.417
2	0.371	5.185
3	-0.549	-7.709
4	0.442	-2.827
5	-0.554	6.921
6	-0.659	0.238
7	0.301	4.252
8	0.373	-11.732
9	0.38	6.044
10	0.495	4.458

Showing 1 to 10 of 22 entries Previous 1 2 3 Next

******

j. DfBetas

Copy CSV PDF Print Search:

	group1	age
1	0.003	0.004
2	0.004	0.004
3	0.004	0.005
4	0.004	0.005
5	0.004	0.005
6	0.006	0.006
7	0.007	0.006
8	0.007	0.007
9	0.007	0.007
10	0.009	0.007

Showing 1 to 10 of 402 entries Previous 1 2 3 4 5 … 41 Next

******

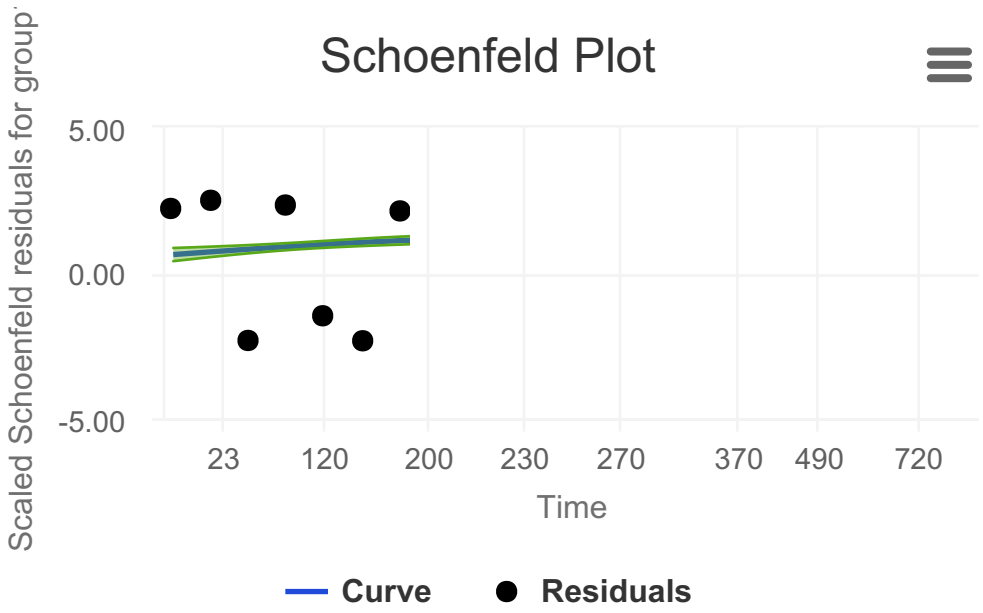
k. Proportional Hazard Test

Copy CSV PDF Print Search:

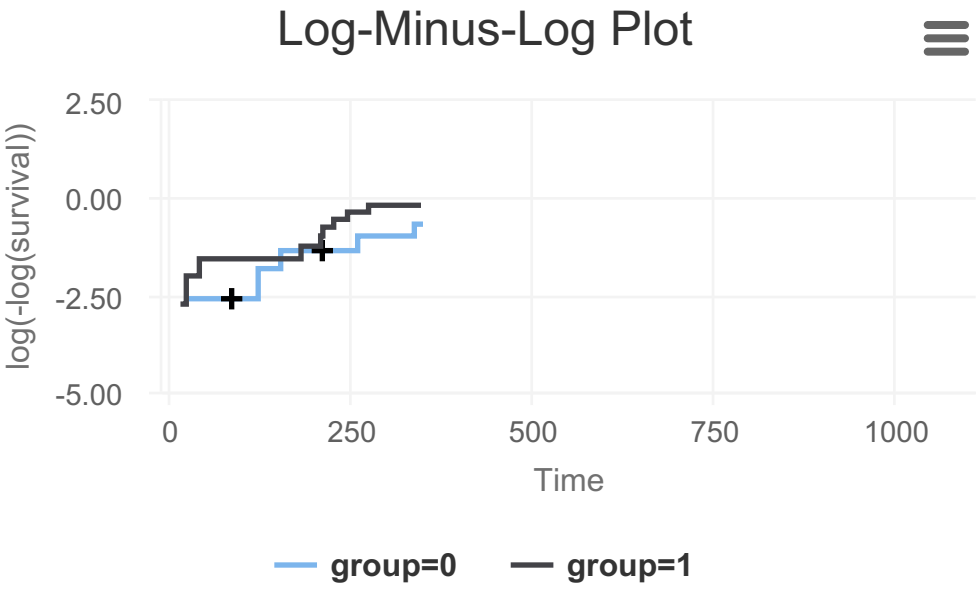
	Rho	Chi-square statistic	p value
group1	-0.257	1.455	0.228
age	-0.176	0.820	0.365
GLOBAL	NA	1.968	0.374

Showing 1 to 3 of 3 entries Previous 1 Next

l. Schoenfeld Plot



l. Log-Minus-Log Plot



2.4. Penalized Cox Regression

Feature selection is an useful strategy to avoid over-fitting, to obtain more reliable predictive results, and to provide more insights into the underlying casual relationships [Penalized feature selection and classification in bioinformatics]. In this section, a feature selection can be made using ridge, elastic net or lasso penalty, especially when there are too many predictors (e.g. `n<<p`).

A Penalized Cox regression analysis can be conducted by applying the following steps:

1. Select the analysis method as `Penalized Cox Regression` from `Analysis` tab.
2. Select suitable variables for the analysis, such as `survival time`, `status variable`
3. If all predictors are continious then one can check the `Select All Variables` option to include all variables in dataset to the feature selection process. If some predictors categorical and others are continious, then uncheck the `Select All Variables` option and select categorical and continuous variables seperately.
4. Define the penalty term using the `Penalty term` slider as follow:

`Penalty term = 0`: ridge penalty `0 < Penalty term < 1`: elastic net penalty `Penalty term = 1`: lasso penalty

5. Select the number of folds for cross-validation. Note that number of folds must be greater than 3.
6. Click `Run` button to run the analysis.

a) Variables in the model

Copy CSV PDF Print Search:

	Variable	Coefficient estimate
1	group	0.298
2	age	0.122

Showing 1 to 2 of 2 entries Previous 1 Next

b) Cross-validation curve

Copy CSV PDF Print Search:

	Variable	Coefficient estimate
1	group	0.298
2	age	0.122

Showing 1 to 2 of 2 entries Previous 1 Next

2.5. Random Survival Forests

A random survival forests analysis can be conducted by applying the following steps:

1. Select the analysis method as `Random Survival Forests` from `Analysis` tab.
2. Select suitable variables for the analysis, such as `survival time`, `status variable`, `category value for status variable`, and categorical and continuous predictors for the model.
3. In advanced options, `interaction terms`, `strata terms` and `time dependent covariates` can be added to the model. Moreover, if there are multiple records for observations, users can specify it by clicking `Multiple ID` checkbox. From `RSF options`, `number of tree`, `bootstrap method`, `randomly selected number of variable`, `minimum number of cases in terminal node`, `maximum depth for a tree`, `splitting rule`, `number of split`, `missing values`, `number of iterations of the missing data algorithm`, `proximity of cases`, `size of bootstrap` and `type of bootstrap` can be adjusted.
4. Click `Run` button to run the analysis.

---

a) Individual Survival Predictions

Copy CSV PDF Print Search:

	14	22	25	40	121	152	180	207	210	225	244	258	273	336
1	1	1	0.957	0.957	0.957	0.957	0.957	0.948	0.948	0.948	0.893	0.893	0.892	0.877
2	0.785	0.445	0.445	0.427	0.0963	0.0939	0.0528	0.0528	0.0204	0.0115	0.0115	0.00674	0.00674	0.00674
3	1	0.996	0.991	0.474	0.472	0.346	0.313	0.311	0.277	0.168	0.158	0.134	0.134	0.118
4	1	1	0.697	0.692	0.692	0.686	0.686	0.685	0.685	0.684	0.679	0.678	0.678	0.583
5	1	1	0.997	0.996	0.996	0.996	0.996	0.93	0.929	0.927	0.731	0.731	0.706	0.705
6	1	1	0.998	0.998	0.998	0.998	0.998	0.988	0.988	0.988	0.987	0.987	0.938	0.937
7	1	1	1	1	1	1	1	0.996	0.996	0.996	0.995	0.995	0.95	0.95
8	1	1	0.916	0.845	0.843	0.79	0.789	0.789	0.788	0.78	0.779	0.765	0.765	0.288
9	1	1	0.947	0.804	0.801	0.718	0.716	0.716	0.715	0.701	0.701	0.667	0.667	0.406
10	0.998	0.992	0.99	0.98	0.959	0.894	0.813	0.813	0.737	0.602	0.602	0.331	0.331	0.312

Showing 1 to 10 of 30 entries Previous 1 2 3 Next

---

b) Individual Survival Predictions OOB

Copy CSV PDF Print Search:

	14	22	25	40	121	152	180	207	210	225	244	258	273	336
1	1	1	0.916	0.916	0.916	0.914	0.914	0.899	0.899	0.899	0.794	0.792	0.792	0.761
2	0.723	0.723	0.723	0.694	0.222	0.217	0.128	0.128	0.0524	0.0308	0.0308	0.018	0.018	0.018
3	1	0.999	0.995	0.995	0.995	0.724	0.691	0.686	0.649	0.436	0.412	0.357	0.355	0.312
4	1	1	1	0.997	0.997	0.989	0.989	0.987	0.987	0.985	0.976	0.974	0.974	0.869
5	1	1	0.994	0.992	0.992	0.992	0.992	0.907	0.905	0.901	0.624	0.624	0.591	0.59
6	1	1	0.996	0.996	0.996	0.996	0.996	0.975	0.975	0.975	0.973	0.973	0.897	0.897
7	1	1	1	1	1	1	1	0.99	0.99	0.99	0.986	0.986	0.89	0.89
8	1	1	0.901	0.708	0.706	0.614	0.611	0.611	0.609	0.593	0.593	0.57	0.57	0.57
9	1	0.999	0.964	0.575	0.569	0.375	0.369	0.369	0.366	0.335	0.335	0.284	0.284	0.187
10	0.993	0.976	0.973	0.95	0.904	0.783	0.627	0.627	0.507	0.347	0.347	0.0193	0.0193	0.00296

Showing 1 to 10 of 30 entries Previous 1 2 3 Next

---

c) Individual Cumulative Hazard Predictions

Copy CSV PDF Print Search:

	14	22	25	40	121	152	180	207	210	225	244	258
1	0	0	0.0427	0.0427	0.0427	0.0434	0.0434	0.0525	0.0525	0.0525	0.11	0.111
2	0.215	0.65	0.65	0.671	1.21	1.21	1.28	1.28	1.36	1.38	1.38	1.39
3	0	0.00425	0.00943	0.533	0.536	0.689	0.738	0.74	0.811	0.986	0.999	1.04
4	0	0	0.303	0.311	0.311	0.319	0.319	0.32	0.32	0.321	0.329	0.33
5	0	0	0.00303	0.00383	0.00383	0.00383	0.00383	0.0707	0.0715	0.074	0.285	0.285
6	0	0	0.00207	0.00207	0.00207	0.00207	0.00207	0.0119	0.0119	0.0119	0.0126	0.0126
7	0	0	0	0	0	0	0	0.00363	0.00363	0.00363	0.00545	0.00545
8	0	0	0.0838	0.159	0.16	0.221	0.223	0.223	0.224	0.237	0.238	0.255
9	0	0.0002	0.0531	0.199	0.203	0.3	0.304	0.304	0.306	0.329	0.329	0.372
10	0.00225	0.00938	0.011	0.0216	0.0465	0.119	0.212	0.212	0.327	0.508	0.508	0.919

Showing 1 to 10 of 30 entries Previous 1 2 3 Next

---

d) Individual Cumulative Hazard Predictions OOB

Copy CSV PDF Print Search:

	14	22	25	40	121	152	180	207	210	225	244	258
1	0	0	0.0844	0.0844	0.0844	0.0865	0.0865	0.102	0.102	0.102	0.214	0.216
2	0.277	0.277	0.277	0.306	0.806	0.814	0.937	0.937	1.11	1.15	1.15	1.18
3	0	0.00057	0.00451	0.00451	0.00522	0.277	0.314	0.319	0.379	0.648	0.677	0.757
4	0	0	0	0.00273	0.00273	0.0115	0.0115	0.0126	0.0133	0.016	0.0255	0.0269
5	0	0	0.00646	0.00808	0.00808	0.00808	0.00808	0.0934	0.0956	0.101	0.409	0.409
6	0	0	0.00443	0.00443	0.00443	0.00443	0.00443	0.0253	0.0253	0.0253	0.0274	0.0274
7	0	0	0	0	0	0	0	0.00966	0.00966	0.00966	0.0145	0.0145
8	0	0	0.099	0.301	0.304	0.414	0.42	0.42	0.423	0.448	0.448	0.476
9	0	0.000545	0.0359	0.434	0.442	0.674	0.686	0.686	0.692	0.745	0.745	0.814
10	0.00666	0.0278	0.0304	0.0543	0.112	0.253	0.44	0.44	0.645	0.897	0.897	1.52

Showing 1 to 10 of 30 entries Previous 1 2 3 Next

e) Error Rate

Copy CSV PDF Print Search:

	Number of tree	Error rate
1	1	0.3
2	2	0.332
3	3	0.367
4	4	0.356
5	5	0.324
6	6	0.318
7	7	0.31
8	8	0.316
9	9	0.313
10	10	0.327

Showing 1 to 10 of 1,000 entries Previous 1 2 3 4 5 … 100 Next

f) Variable Importance

Copy CSV PDF Print Search:

	variableImp
group	-0.00251
age	0.181

Showing 1 to 2 of 2 entries Previous 1 Next

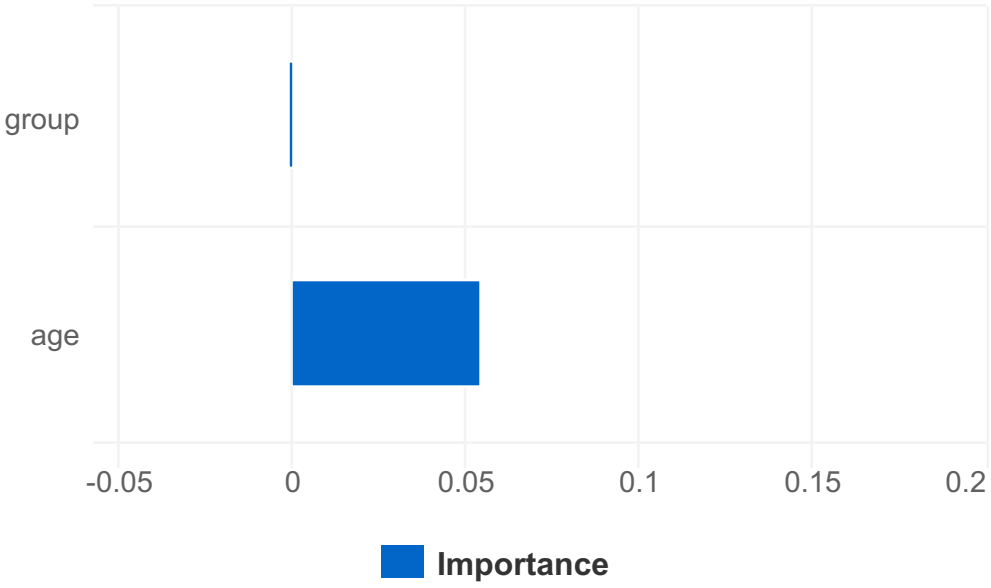
f) Variable Importance

Copy CSV PDF Print Search:

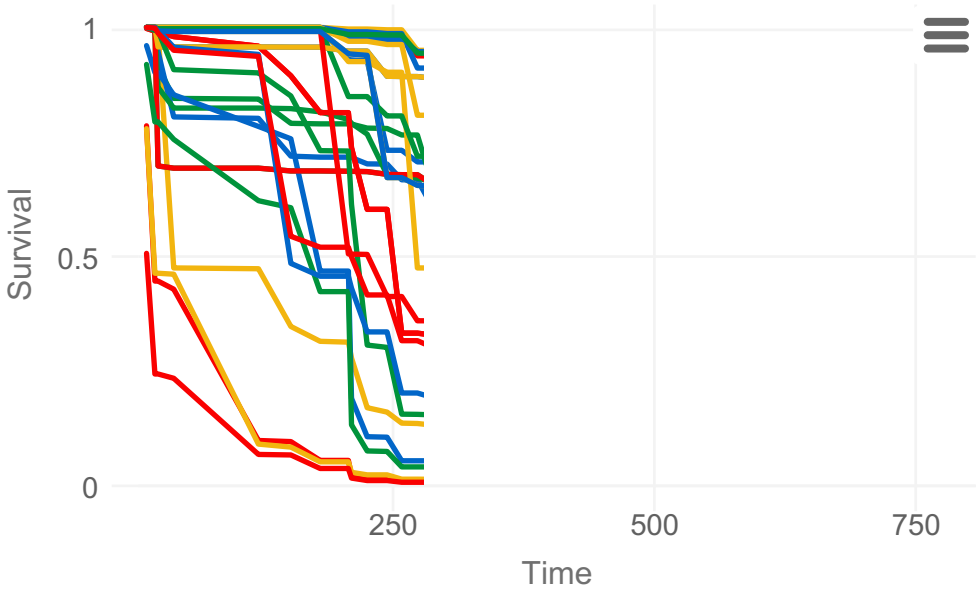
	variableImp
group	-0.00251
age	0.181

Showing 1 to 2 of 2 entries Previous 1 Next

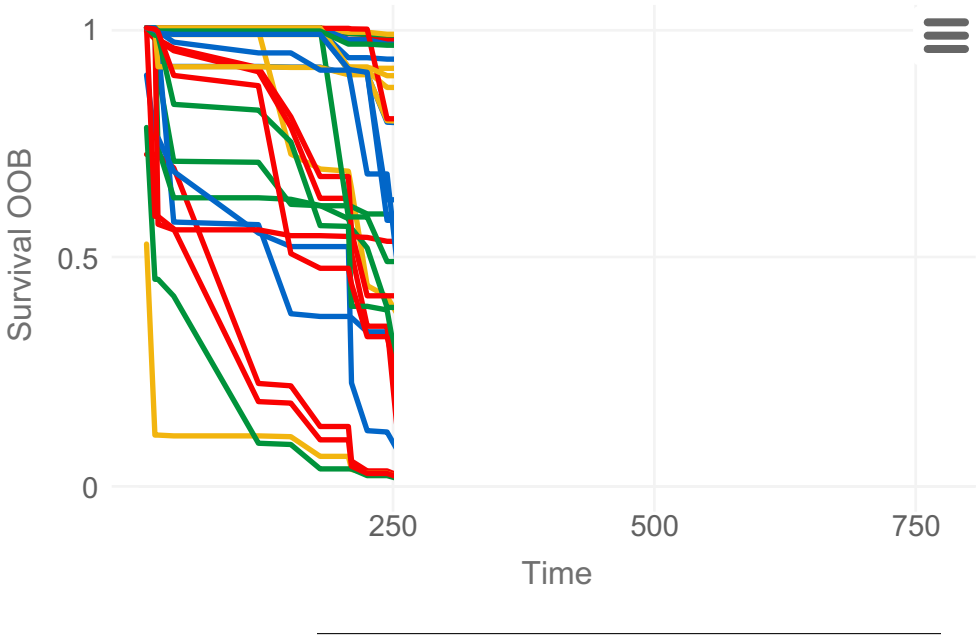
g) Variable Importance Plot



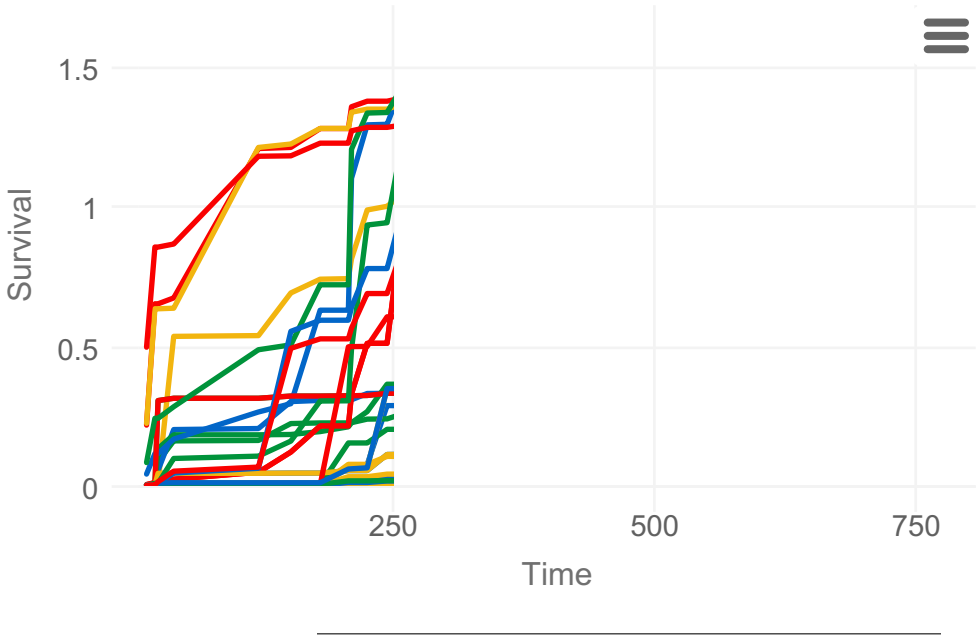
g) Survival Plot



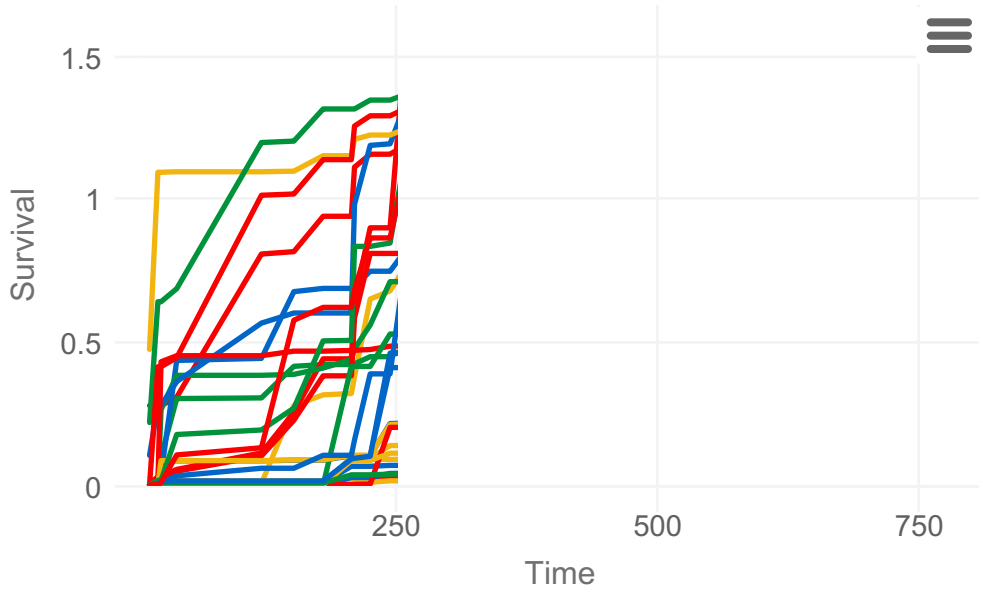
h) Survival OOB Plot



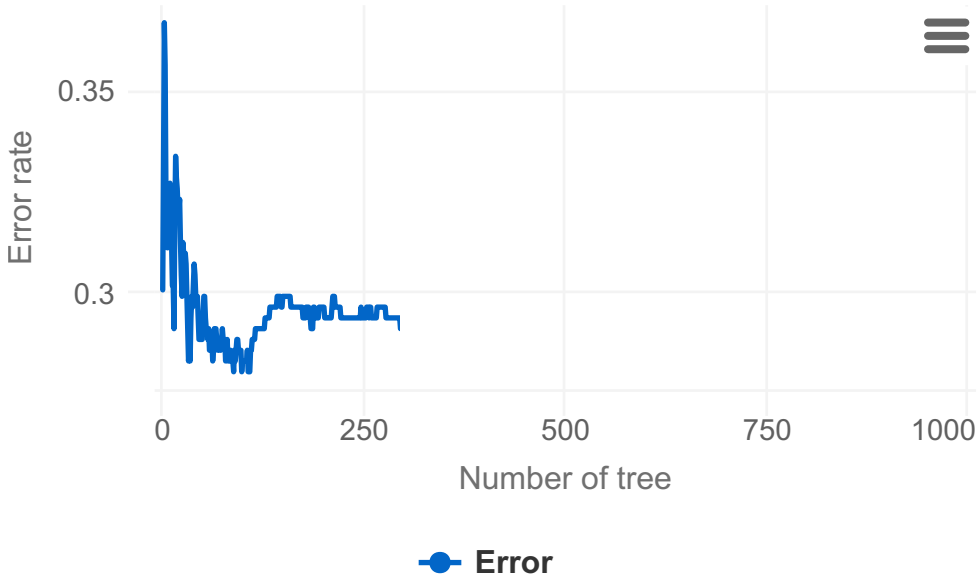
h) Hazard Plot



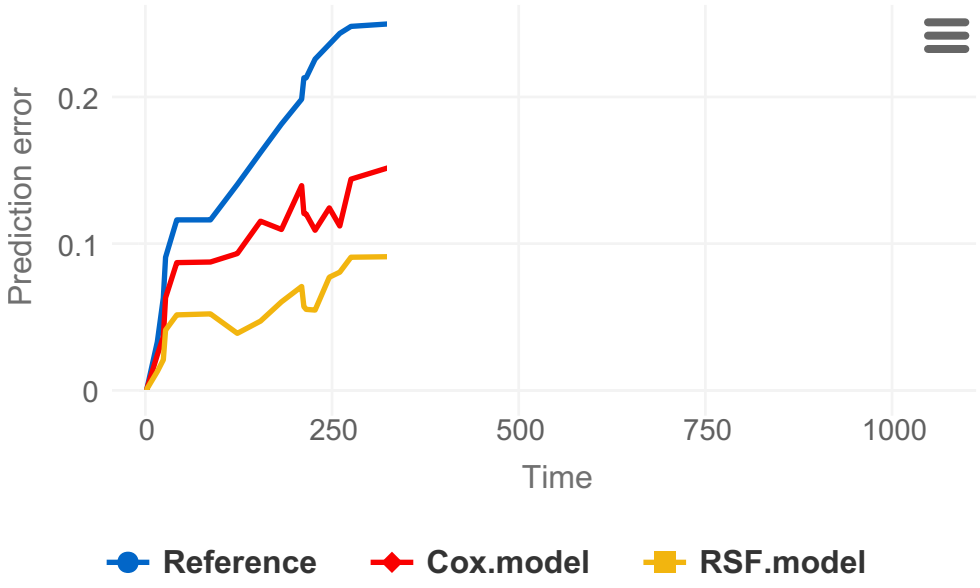
h) Hazard OOB Plot



i) Error Rate Plot



i) Cox vs RSF


Prediction error
0.2
0.1
0
0
250
500
750
1000
Time
Reference
Cox.model
RSF.model