

Model Evaluation

Considerations for Time-to-Event Studies

Daniel Saggau

11/15/2020

Overview

- ▶ Time to Event Studies:

What is a time to event study ?

- ▶ Classical Model Evaluation:

Why cant we use them?

- ▶ TTS Model Evaluation:

How to derive these methods (c-index, brier)?

- ▶ Discussion :

What are the shortcomings of these methods?

- ▶ Further Considerations:

What solutions exist?

Time-to Event Studies

- ▶ Analysis working with (right) censored data
- ▶ Highly relevant for clinicians in the field of medical statistics
e.g. looking at when a patient dies or when he gets a disease
(clinical/epidemiological studies)
- ▶ In Economics/Finance e.g. to examine when a subject/borrower
will default or when a subject will find/lose a job
- ▶ Operations research to predict the time a machine will break

Basic Notations & Concepts

- ▶ Time T and Survival S
- ▶ From hazard to cumulative hazard to survival
- ▶ Hazard $h(t,x)$ is the eminent probability of death a specific point in time
- ▶ Capital H is the cumulative hazard
- ▶ non-parametric hazard models (KM) vs. semi-parametric proportional hazard model

Model Evaluation - Considerations

(1) *What type of study are we dealing with?*

Diagnostic vs. Prognostic Study

(2) *What are the components of our model evaluation metric?*

Discrimination: Are we able to correctly discriminate between e.g. sick and healthy patients ? **Calibration:** How concise is our prediction accuracy ? **Clinical Usefulness:** Will our model create more benefits than harm?

Classical Model Evaluation Tools for Classification Tasks

Working with *Label* vs. working with *Probability*

- ▶ Brier Score (probability from true class label)
- ▶
- ▶ AUC/ROC (receiver operating characteristics)
- ▶ Mis-classification Error rate (rate of incorrect classification)
- ▶ ACC (rate of correct classifications)
- ▶ Concordance Statistics

Brier Score

- Based on loss function

MSE for Regression (L2 Loss):

$$BS = \frac{1}{n} \sum_{i=1}^n (y^{(i)} - \hat{y}^{(i)})^2$$

Where: the $MSE \in [0; \infty)$

The Brier Score is the MSE for Classification:

$$BS = \frac{1}{n} \sum_{i=1}^n (\hat{\pi}(x^{(i)}) - y^{(i)})^2$$

The general version of the brier score looks at a specific point in time

Confusion Matrix

Sensitivity:

- ▶ deals with values above the threshold among the subject group which do endure an event
- ▶ Another common name for Sensitivity is the true positive rate.

$$TPF = \frac{TP}{TP+FN}$$

Specificity:

- ▶ deals with false negatives, hence patients with a disease we classify as not having any diseases
- ▶ Another name for specificity is the true negative rate

$$TNR = \frac{TN}{TN+FP}$$

Why cant we use traditional model evaluation tools for time to event studies?

- ▶ Working with censored data “
- ▶ Right censored data (event after follow up) vs. left censored data (event was not recorded when it occurred initially)
- ▶ We need to estimate survival of patients without having data on e.g. death
- ▶ Also, we need to provide measure over time

Early approaches: - excluding subjects with right censored data and only evaluate on the complete data

From ROC to C-Statistic to C-index

- ▶ Advancement of ROC/AUC
- ▶ Further modification leads to c-index
- ▶ concordance pairs divided
- ▶ concordance $\Rightarrow x_1 > x_2 \text{ -- } y_1 > y_2$
- ▶ Harell's C

c-index

- ▶ studying concordance pairs of subjects
- ▶ addressing right censored data via inverse of the probability of censoring weighted estimate (of concordance probability)
- ▶ kendall's tau as conservative weight
- ▶ Summary measure (over all time) based on the AUC

$$C - index = \frac{\Delta_j \times \sum_{i,j} 1_{T_i > T_j} \times 1_{\eta_i > \eta_j}}{\Delta_j \times \sum_{i,j} 1_{T_i > T_j}}$$

- ▶ Where 1 are indicator-functions:

```
##
```

```
## randomForestSRC 2.9.3
```

```
##
```

```
## Type rfsrc.news() to see new features, changes, and bug
```

```
##
```

```
##
```

```
## The c-index for right censored event times
```

```
##
```

```
## Prediction models:
```

IBS

- ▶ called cumulative predictive error curves == continuous ranked probability score (crps)
- ▶ area under the prediction error curve
- ▶ Integral over all points in time to get one summary value henceforth called “integrated” BS
- ▶ able to build a R^2 like measure where we divide MSE of a model with a different MSE of reference model
- ▶ Where L is a loss function of the S (the probability that the event of interest has not taken place yet) and time
- ▶ t is the time of the event (death) and t^* the time before death
- ▶ $G(t)$ is the $P(C > t)$, so where the censored time is longer than the time (in `mlr3proba` via `survfit` == KM Estimate)
- ▶ When selecting integrated == FALSE then we looking at specific time

For the population mean:

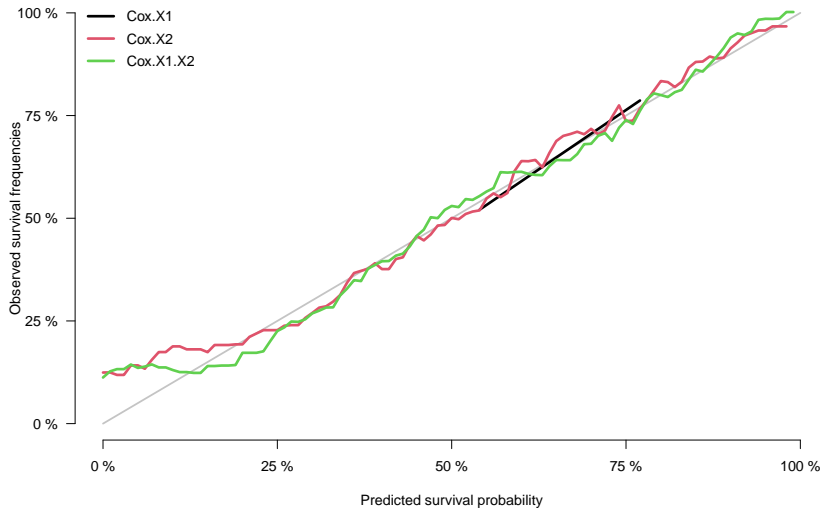
$$L(S, t|t^*) = \frac{1}{N} \sum_{i=1}^N L(S_i, t_i|t^*) \quad (9)$$

Mean Population:

$$L(S, t|t^*) = \frac{1}{NT} \sum_{i=1}^N \sum_{j=1}^T L(S_i, t_i|t^*)$$

- ▶ N = Number of observations
- ▶ S_i is the predicted survival function

Calibration Plot



Summary Prediction Error Curves

##

Prediction error curves

##

##

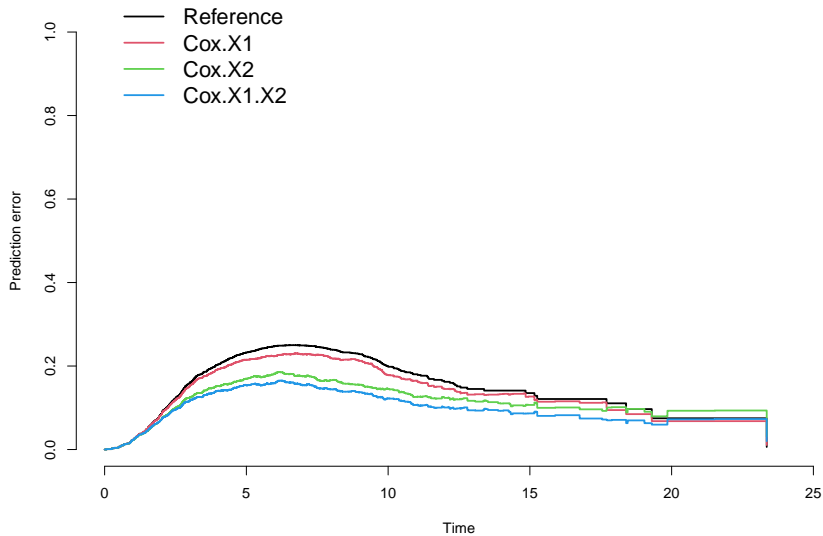
No data splitting: either apparent or independent test s

##

AppErr

| ## | time | n.risk | Reference | Cox.X1 | Cox.X2 | Cox.X1.X2 |
|------|------|--------|-----------|--------|--------|-----------|
| ## 1 | 0 | 1000 | 0.000 | 0.000 | 0.000 | 0.000 |
| ## 2 | 5 | 471 | 0.233 | 0.215 | 0.170 | 0.154 |
| ## 3 | 10 | 105 | 0.199 | 0.178 | 0.145 | 0.122 |
| ## 4 | 15 | 17 | 0.135 | 0.127 | 0.107 | 0.087 |
| ## 5 | 20 | 2 | 0.075 | 0.068 | 0.093 | 0.072 |

Plotting prediction error



Cumulative Prediction Error

##

Integrated Brier score (crps):

##

IBS[0;time=0) IBS[0;time=5) IBS[0;time=10) IBS[0;time=15)

Reference 0 0.117 0.177

Cox.X1 0 0.111 0.164

Cox.X2 0 0.091 0.129

Cox.X1.X2 0 0.085 0.116

IBS[0;time=20)

Reference 0.156

Cox.X1 0.144

Cox.X2 0.119

Cox.X1.X2 0.101

##

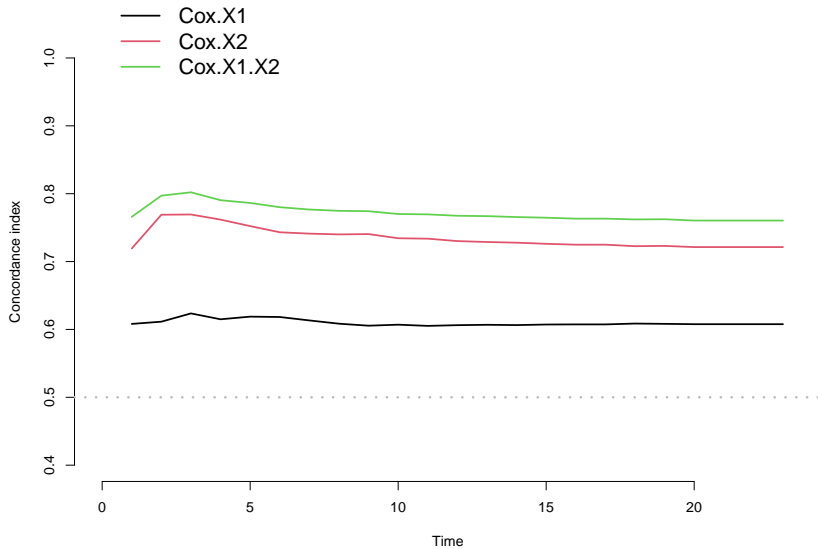
Integrated Brier score (crps):

##

IBS[0;time=23.4)

Reference 0.144

c-index plot



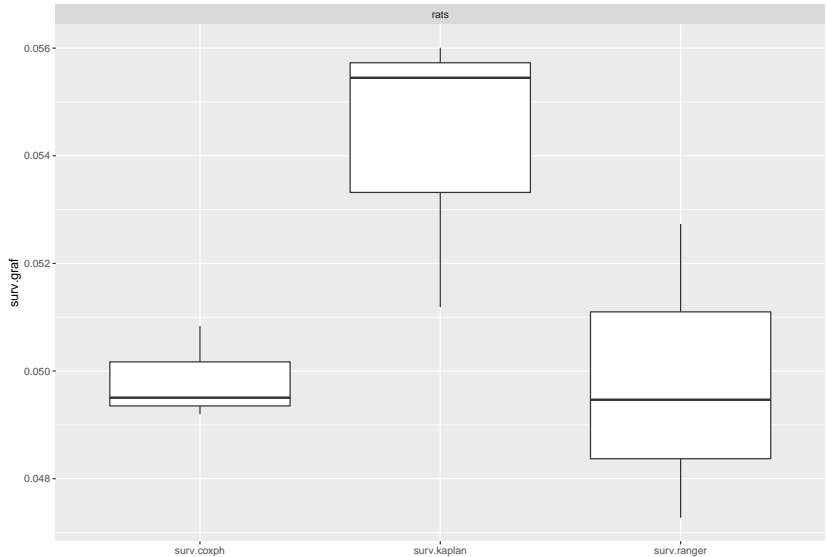
mlr3Proba

- ▶ van Houwelingen's Alpha Calibration
- ▶ van Houwelingen's Beta Calibration
- ▶ Integrated Graf Score (other Name for IBS based on Author Graf)
- ▶ Integrated Log Loss
- ▶ Log Loss

Further measures via survAUC package:

- ▶ Uno's AUC/TPR/TNR
- ▶ Song and Zhou's AUC/TNR/TPR
- ▶ Chambless and Diao's AUC
- ▶ Hung and Chiang's AUC

mlr3Proba Example



Discussion

- ▶ Integrated Brier Score accounts for both calibration and discrimination
- ▶ Irrespective, neither model accounts and leaves room for improvement

Literature and Recommendations

Introduction:

- ▶ Steyerberg, E. W., Vickers, A. J., Cook, N. R., Gerds, T., Gonen, M., Obuchowski, N., ... & Kattan, M. W. (2010). Assessing the performance of prediction models: a framework for some traditional and novel measures. *Epidemiology* (Cambridge, Mass.), 21(1), 128.

Comparative Study:

- ▶ Kattan, M. W., & Gerds, T. A. (2018). The index of prediction accuracy: an intuitive measure useful for evaluating risk prediction models. *Diagnostic and prognostic research*, 2(1), 7.

Use Cases:

https://rpubs.com/kaz_yos/survival-auc

<https://datascienceplus.com/time-dependent-roc-for-survival-prediction-models-in-r/> <https://rdr.io/cran/pec/>

<https://adibender.github.io/pammttools/>