Model Evaluation

Considerations for Time-to-Event Studies

Daniel Saggau

11/15/2020

Overview

- Time to Event Data
- Terminology
- Classical Model Evaluation Tools
- Integrated Brier Score
- Concordance-Index
- Discussion
- Further Considerations

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 structure

- Time T and Survival S
- 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
- From hazard to cumulative hazard to survival
- Survival Probability

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

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

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

Brier Score - Methods

ROC/AUC/Concordance Statistics - Methods

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 occured intially)
- 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 == x1>x2-> y1>y2
- Harell's C

c-index

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

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

- Where 1 are indicator-functions:

mlr3 Proba Applications:

- van Houwelingen's Alpha Calibration
- van Houwelingen's Beta Calibration
- Integrated Graf Score
- 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

```
##
##  randomForestSRC 2.9.3
##
##  Type rfsrc.news() to see new features, changes, and bug
##

##
```

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 Rˆ2 like measure where we divide MSE of a model with a different MSE of reference model

For the Individual:

$$L(S, t|t^*) = [(S(t^*)^2)I(t) \leq t^*, \delta = 1)(\frac{1}{G(t)})]$$

$$+[((1 - S(t^*))^2)I(t > t^*)(\frac{1}{G(t^*)})]$$

- Where L is a loss function of the S(the probability that the event of interest has not taken place yet) and time

For the population mean:

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

Mean Population:

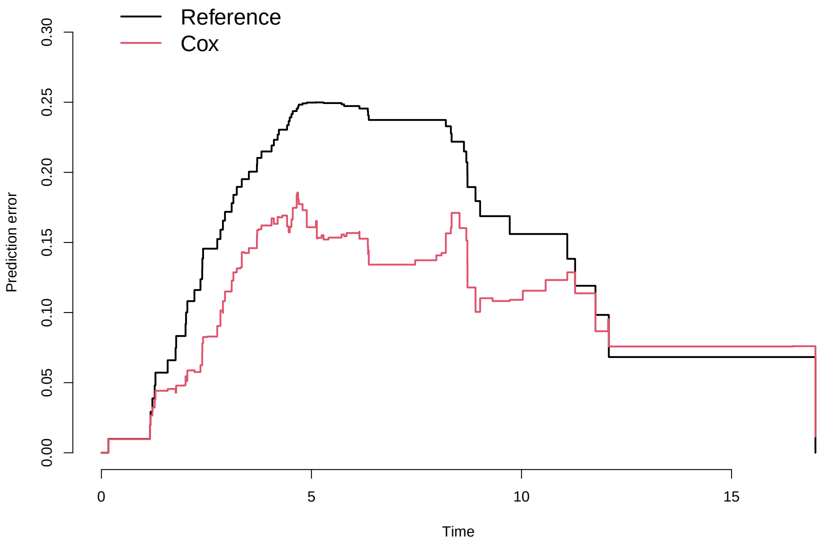
$$L(S, t|t^*) = \frac{1}{NT} \sum_{i=1}^{N} \sum_{j=1}^{T} L(S_i, t_i|t^*)$$

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

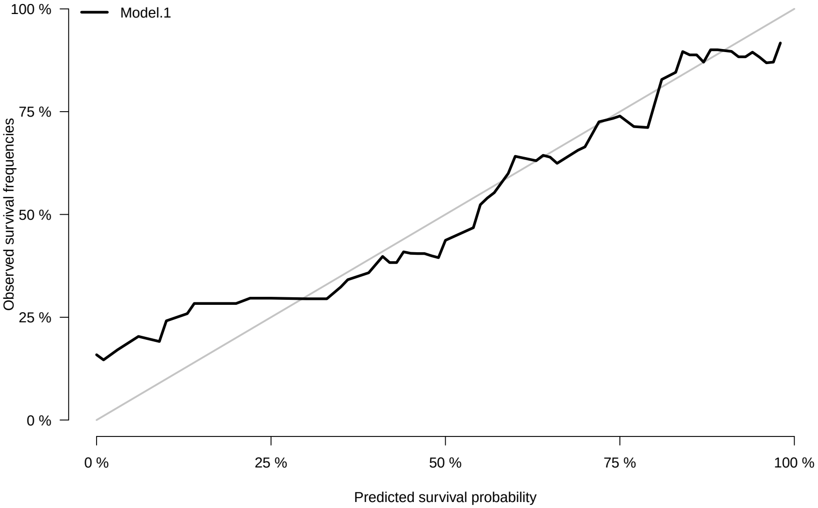
Prediction Error Curve Based on

```
## No covariates  specified: Kaplan-Meier for censoring tin
##
## Integrated Brier score (crps):
##
##           IBS[0;time=0) IBS[0;time=0.25) IBS[0;time=0.5)
## Reference             0            0.003           0.007
## Cox                   0            0.003           0.006
##           IBS[0;time=1)
## Reference         0.008
## Cox               0.008
```

Prediction Error Curve



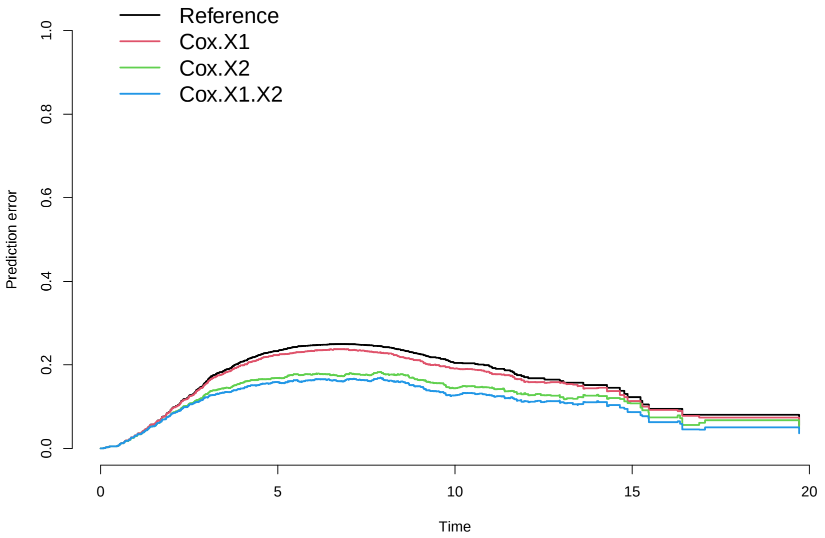
Calibration Plot



Summary Prediction Error Curves

```
##
## Prediction error curves
##
##
## No data splitting: either apparent or independent test s
## Warning in summary.pec(PredError): Missing times argumen
##
##  AppErr
##     time n.risk Reference Cox.X1 Cox.X2 Cox.X1.X2
## 1  0.000   1000     0.000  0.000  0.000     0.000
## 2  2.907    750     0.153  0.150  0.126     0.120
## 3  4.989    500     0.233  0.223  0.169     0.159
## 4  7.738    250     0.246  0.230  0.181     0.166
## 5 21.394      0     0.003  0.006  0.015     0.010
```

Plotting prediction error



Discussion

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

Literature

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.