Discrimination and Calibration

Terry Therneau

February 28, 2022

1 Introduction

This vignette is in progress and not yet complete.

2 Discrimination

2.1 Pseudo R^2 measures

There have been many attempts to define an overall "goodness of fit" criteria for the Cox model which would be parallel to the widely used R^2 of linear models. A direct analog is hampered by the issues of censoring and scale. Censoring is the major technical impediment: how do we define error for an observation known only to be $> t_i$? A potentially larger issue is that of scale: if we have two $(t, \hat{t})$ pairs of (3 m, 6 m) and (9 yr, 10 yr), which one represents the greater error? On an absolute scale the second is the larger difference, but in a clinical study the first may be more important. Simon xxx has a nice discussion of the issue.

Rather than a direct extension of R^2, subject to the issues raised above, the most common approach has been based on pseudo-R^2 measures. These re-write the linear model statistic in another way, and then evaluate the alternate formula.

$$\begin{aligned} R^2 &= 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \overline{y})^2} \\ &= 1 - \left(\frac{LL(\text{intercept})}{LL(\text{full})}\right)^{2/n} \qquad (1) \\ &= \frac{\text{var}(\hat{y})}{\text{var}(\hat{y}) + \sigma^2} \qquad (2) \end{aligned}$$

Equation (1) is the Cox and Snell formula, for a Cox model replace the linear model log-likelihood (LL) with the partial likelihood of the null and fitted models. This gives the measure proposed by Nagelkerke [**?**] which was part of the standard printout of `coxph` for many years. It has, however, been recognized as overly sensitive to censoring.

The measure of Kent and O'Quigley [**?**] is based on (2). Replace $\hat{y}$ with the Cox model linear predictor $\eta = X\hat{\beta}$ and $sigma^2$ by $\pi^2/6$. The latter is based on an extreme value distribution, and the equivalence of the Cox model to a transformation model.

Royston and Sauerbrei replace the risk scores η with a normal-scores transform

$$\begin{aligned} s_i &= \Phi^{-1}\left(\frac{r_i - 3/8}{n + 1/4}\right) \\ r_i &= \text{rank}(\eta_i) \end{aligned}$$

then re-fit the Cox model using s as the single covariate. Since $\text{var}(s) = 1$ by design, the variance of the normalized risk score will be captured by the coefficient β^2 from the re-fit; while the Cox model estimate of the variance of β is used as an estimate of variance. They then further define a measure $D = \beta\sqrt{8/\pi}$. The rationale is that $E(s|s > 0) = \sqrt(2/\pi)$, so D represents the hazard ratio between a random draw from the bottom half of the s distribution to a random draw from the top half. This value is then 'comparable' to the hazard ratio for a binomial covariate such as treatment. (We prefer to use the 25th and 75th quantiles of the risk score, untransformed, for this purpose, i.e. the "middle of the top half" versus the "middle of the bottom half", rather than mean(bottom) vs. mean(top).)

A issue with the Royston and Sauerbrei approach is that although the risk scores from a fitted Cox model will sometimes be approximately symmetric, there is no reason to assume that this should be so. In medical data the risks are often right skewed: it is easier to have an extremely high risk of death than an extremely low one (there are no immortals). For the well known PBC data set, for instance, whose risk score has been validated in several independent studies, the median-centered risk scores range from -2 to 5.4. Remember that even in the classic linear model, $\hat\beta$ and the residuals are assumed to Gaussian, but no such assumption is needed for y or $\hat y$; in fact such a distribution is uncommon. See "Health, Normality and the Ghost of Gauss" [?] for a good discussion of this topic.

Göen and Heller [?] create a pseudo-concordance that also uses only the risk scores. It is based on the fact that, if proportional hazards holds, then the time to event t_i and t_j for two subjects satisfies

$$P(t_i > t_j) = \frac{1}{1 + \exp(\eta_j - \eta_i)} \tag{3}$$

They then propose the estimate

$$C_{GH} = \frac{2}{n(n-1)} \sum_{\eta_i > \eta_j} \frac{1}{1 + \exp(\eta_j - \eta_i)} \tag{4}$$

which can be translated from the (0,1) concordance scale to a (-1, 1) range via $R^2_{GH} = 2C - 1$, if desired. If there is no relationship between X and t then $\beta = 0$ and $\eta = 0$, leading to a concordance of $1/2$. A claimed advantage of the GH measure over the usual concordance is that it is not affected by censoring. A primary disadvantage is that it is based on the assumption that the model is completely correct; the assumption that proportional hazards holds for all time, even well beyond any observed data, is particularly unlikely.

All of the above measures are computed by the `royston` command.

Evaluation of new data

When applying these measures to new data, for an existing model, it is necessary to first compute a scaling factor. That is, compute the vector of risk scores $\eta = X\beta$ using the coefficients of the prior model and covariates from the new data, and then fit a new Cox model using the response from the new data with η as the only predictor. The rescaled risk scores $\hat\beta\eta$ are then used in the formulas.

To see why rescaling is necessary, assume a case where the validation data set was an exact copy of the development data set, but with an error: at some point the survival time column had been randomly re-ordered but without perturbing the other columns. The survival time is now unrelated to the linear predictor η, yet the values of the pseudo R-squared and C statistic would be unchanged from the original fit, since η is unchanged. In this extreme case the rescaling fit would have coefficient 0, leading to the appropriate conclusions. Rescaling is done automatically by the `royston` command when the call includes a `newdata` argument.

2.2 Concordance

The most commonly used measure of discrimination is the concordance. [merge with concordance vignette]

3 Calibration

4 Computing IPC weights

4.1 Ties

The computation of inverse of probability of censoring (IPC) weights would appear to be a straightforward task, but has some tricky edge cases. The simple algorthm, used by many, is to do an ordinary Kaplan-Meier with censoring as the endpoint, what is somtetimes called the reverse KM, and then simply read off numbers from the graph, i.e., use the following bit of code in the kernel.

```
cfit <- survfit(Surv(time, 1-status) ~ 1, data=mydata)
cprob <- summary(cfit, times= mydata$time)$surv)
```

There are two problems with this. The first is that the Kaplan-Meier creates a right continuous estimate, and IPC weights need to be left continuous. That is, if we want a censoring weight at time t, this should be based on the probability of being censored *before* time t; $C(t-)$. The `summary` function returns $C(t)$. The second is that if the starting data set includes any time points with both an event and a censoring at that time point, the observation with an event is *not* at risk for censoring at that time, but in a reverse KM they will be counted. Complicating things even further is the issue of round-off error as dicussed in the vignette on tied times.

A general solution to the above is shift all of the censoring times right by a small amount ϵ, before using `survfit` to compute $C(t)$. This resolves the issue of tied death and censoring times. When time values are looked up on the resulting curve, the result will effectively be based on the right-continuous version of $C(t)$. Care much be used to choose ϵ sufficiently small so that the shifted values do not overstep another unique time point, and sufficiently large that the shift is not lost due to roundoff error (in floating point arithmetic 1000 + 1e-16 =1000). Then, at any chosen cutoff τ, the IPW weights will be 0 for any observation censored before τ, and $1/C(\min(t_i, \tau))$ otherwise, where t_i is the follow-up time for observation i.

There are two fairly simple checks on any implementation of IPW. The first is to compute weights using the maximum uncensored time as the cutoff τ, then the weighted emprical CDF of t should exactly agree with the Kaplan-Meier. The second is that the sum of the weights must equal n, the number of observations, for any cutoff τ. Ignoring the issue of tied event and censoring times creates weights that are too small, and using right-continuous vs. left continuous C creates weights that are too large.

4.2 Counting process data

Modification of the redistribute to the right (RTTR) algorithm for counting data is straightforward. First, only actual censorings cause a redistribution. For example, assume we have a data set with time-dependent covariates, and a subject with three (time1, time2, status) intervals of (0, 5, 0), (5,18, 0) and (18, 25, 0). Only the last of these, time 25, is an actual censoring time. For counting process data the `rttright` function will insist on an `id` statement to correctly group rows for a subject, and makes use of `survcheck` to insure that there are no gaps or overlaps. (The `survfit` routine imposes the same restriction).

A general form of the RTTR algorithm that allows for multiple states has the following features

1. Censoring weights $c_i(t)$ for each observation are explicitly a function of time, and sum to 1.
2. When an observation is censored, the weight it redistributied to all other observations in the same state.
3. At any time, for any given state, the non-zero weights for all observations in that state are proportional to prior case weights w_i.

The estimated probability in state j is estimated as a sum of weights of those currently in state j.

$$p_j(t) = \sum_{s_i=j} c_i(t)$$

This estimate agrees with the Aalen-Johansen estimate.

Point 3 means that whenenever an observation changes state, this necessates re-balancing of the weights in the new state. As an example assume a model with three states 1 → 2 → 3, and 5 subjects who start in state 1, and the following data set where a state of 0 = censoring. At time 2.1 the weights are 1/5, 0, 4/15, 4/15, 4/15, with observation 1 in state 2 and 3–5 in state 1, and at time 3 observation 3 transitions to state 2. At time 4, observation 1 transitions to state 3 with a weight of (1/5 + 4/15)/2.

```
  id time1 time2 state
1  1     0     1     2
2  1     1     4     3
3  1     4     8     0
4  2     0     2     0
5  3     0     3     2
6  3     3     6     0
7  4     0     7     0
8  5     0     8     0
```

For an absorbing state, i.e., one with no departures, the rebalancing step is not technically necessary since it does not change the sum. As a practical matter in the code, there is actually no need to rebalance until there is a departure of an observation to another state. This fits well with a design decision of the survival package to not force declaration of the aborbing states by the user, but to simply notice them as those states where no one departs.

For the case of a simple alive-dead or competing risks model, the weights generated by the RTTR approach also have a simple representation as $1/G(t)$ for some censoring distribution G. The keys that makes it work for these two cases are that everyone starts in the same state and that subjects are only censored from that state. There is in essence only one censoring distribution to keep track of. There is not natural way (that we have seen) to reflect RTTR weights that involve rebalancing as the inverse of a censoring distribution G.