Concordance

Terry Therneau, Elizabeth Atkinson

November 26, 2018

1 The concordance statistic

Use of the concordance statistic for Cox models was popularized by Harrell [1], and it is now the most used measure of goodness-of-fit in survival models. In general let y_i and x_i be observed and predicted data values, in the most common case $x = \hat{\eta}$, the linear predictor from a fitted model. The concordance is defined $P(x_i > x_j | y_i > y_j)$, the probability that the prediction x goes in the same direction as the actual data y. A pair of obervations i, j is considered concordant if the prediction and the data go in the same direction, i.e., $(y_i > y_j, x_i > x_j)$ or $(y_i < y_j, x_i < x_j)$. The concordance is the fraction of concordant pairs. For a Cox model remember that the predicted survival $\hat{y}$ is longer if the risk score $X\beta$ is lower, so we have to flip the definitions of condordant and discordant. For now, ignore this detail and use the usual definition for exposition.

One wrinkle is what to do with ties in either y or x. Such pairs can be ignored in the count (treated as incomparable), treated as discordant, or given a score of 1/2. Let C, D, T_x, T_y and T_{xy} be a count of the pairs that are concordant, discordant, and tied on x (but not y), tied on y (but not x), and tied on both. Then

$$\tau_a = \frac{C - D}{C + D + T_x + T_y + T_{xy}} \tag{1}$$

$$\tau_b = \frac{C - D}{\sqrt{(C + D + T_x)(C + D + T_y)}} \tag{2}$$

$$\gamma = \frac{C - D}{C + D} \tag{3}$$

$$d = \frac{C - D}{C + D + T_x} \tag{4}$$

- Kendall's tau-a (1) is the most conservative; essentially treating ties as failures
- The Goodman-Kruskal γ statistic (3) ignores ties in either y or x.
- Somers' d (4) treats ties in y as incomparable; pairs that are tied in x (but not y) score as 1/2. The AUC from logistic regression is equal to Somers' d.

All three of the above range from -1 to 1. The concordance is $(d + 1)/2$.

The concordance has a natural interpretation as an experiment: present pairs of subjects one at a time to the physician, statistical model, or some other oracle, and count the number of

correct predictions. Pairs that have the same outcome are not put forward for scoring (that would be unfair); and if the oracle cannot decide it makes a random choice. This leads to $C + T_x/2$ correct selections out of $C+D+T_x$ choices, which is easily seen to be equal to $(d+1)/2$. Kendall's tau-b has a denominator of the same type, but treats x and y symmetrically.

This hypothetical experiment gives a baseline insight into the concordance. A value of 1/2 corresponds to using a random guess for each subject, and values of $< .55$ are not very impressive. The ordering for some pairs of subjects will be obvious, and someone with almost no medical knowledge could do nearly that well by marking those pairs and using a coin flip for the rest. Values of less than 1/2 are possible - some stock market analysts come to mind.

For survival data any pairs which cannot be ranked with certainty are also considered incomparable. For instance y_i is censored at time 10 and y_j is an event (or censor) at time 20. Subject i may or may not survive longer than subject j, and so it is not possible to tell if the rule has ranked them correctly or not. Note that if y_i is censored at time 10 and y_j is an event at time 10 then $y_i > y_j$. For stratified models, observations that are in different strata are also considered to be incomparable.

2 Examples

The `concordance` function available in the survival package can be used to estimate concordance with various types of models including logistic and linear regression.

```
> # logistic regression using Fisher's iris data
> fit1 <- glm(Species=="versicolor" ~ ., family=binomial, data=iris)
> concordance(fit1)  # this gives the AUC
Call:
concordance.lm(object = fit1)

n= 150
Concordance= 0.8258 se= 0.03279
concordant discordant     tied.x     tied.y    tied.xy
      4129        871          0       6174          1
> # linear regression
> fit2 <- lm(karno ~ age + trt, data=veteran)
> concordance(fit2)  # 2*concordance-1 = somers' d
Call:
concordance.lm(object = fit2)

n= 137
Concordance= 0.5397 se= 0.032
concordant discordant     tied.x     tied.y    tied.xy
      4322       3679         90       1211         14
```

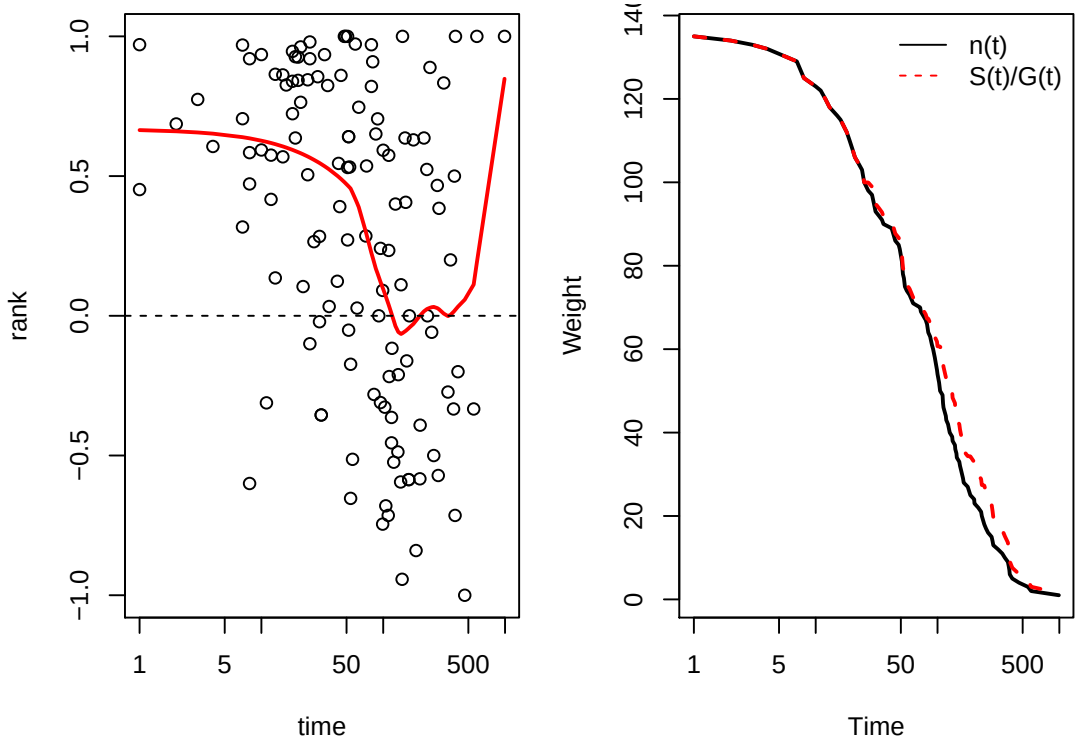
Primarily the focus of this vignette will be on survival data.

```
> # parametric survival regression
> fit3 <- survreg(Surv(time, status) ~ karno + age + trt, data=veteran)
> concordance(fit3)
Call:
concordance.survreg(object = fit3)

n= 137
Concordance= 0.7122 se= 0.02232
concordant discordant     tied.x     tied.y    tied.xy
      6263       2527         14         39          0
> # 3 Cox models
> fit4 <- coxph(Surv(time, status) ~ karno + age + trt, data=veteran)
> fit5 <- update(fit4, . ~ . + celltype)
> fit6 <- update(fit5, . ~ . + prior)
> ctest <- concordance(fit4, fit5, fit6)
> ctest
Call:
concordance.coxph(object = fit4, fit5, fit6)

n= 0
     concordance     se
fit4      0.7119 0.0224
fit5      0.7384 0.0210
fit6      0.7359 0.0212

     discordant concordant tied.x tied.y tied.xy
fit4       6261       2529     14     39       0
fit5       6499       2301      4     39       0
fit6       6478       2324      2     39       0
```

As shown in the last example, the concordance for multiple fits can be obtained from a single call. The variance-covariance matrix for all three concordance values can be obtained using `vcov(ctest)`; this is used in an example below to formally test equality of two concordance values. The above also shows that addition of another variable to a fitted model can decrease the concordance.

Concordance can be visualized using its contribution at each event time as shown below. The y value on the graph is a rank of each subject's risk score as compared to the risk scores of all those subjects with longer survival, r = fraction with lower scores - fraction with higher scores. In a highly predictive Cox model the subject with the highest risks will die soonest. A log transformation on the x-axis was used to visually spread out the points. Somers' d is the weighted average of these ranks and the concordance is $c = (d + 1)/2$. After about 50 days, the baseline data has little predictive ability and the average ranks are close to zero. The second panel shows two possible weights, based on the number of subjects available at a given time, where $n(t)$, the default choice, is the number of comparable pairs at each time point. Alternate weights such as S/G are discussed in section 6. For both choices the weights decrease precipitously over time and the final average is based largely on the left hand portion of the plot.

```
> par(mfrow=c(1,2))
> c3 <- concordance(fit3, ranks=TRUE)
> c4 <- concordance(fit3, ranks=TRUE, timewt="S/G")
> plot(rank ~ time, data=c3$ranks, log='x')
> with(c3$ranks, lines(lowess(time, rank), col=2, lwd=2))
> abline(0,0, lty=2)
> matplot(c3$ranks$time, cbind(c3$ranks$timewt, c4$ranks$timewt),
          type="l", col=c("black", "red"), lwd=2,
          xlab="Time", ylab="Weight", log="x")
> legend("topright", legend=c("n(t)","S(t)/G(t)"), lty=1:2,
         col=c("black","red"), bty="n")
```



3 Connection to the proportional hazards model

Watson and Therneau [6] show that the numerator of Somers' d for a response y and predictor x can be re-written as

$$C - D = 2 \sum \delta_i n(t_i) \left[r_i(t_i) - 1/2 \right] \tag{5}$$

where C and D are the total number of concordance and discordant pairs, $n(t)$ is the number of subjects still at risk at time t, and $r_i(t)$ is the rank of x_i among all those still at risk at time t, where the rank is defined such that $0 \leq r \leq 1$. It turns out that equation (5) is exactly the score statistic for a Cox model with a single time-dependent covariate $n(t)r(t)$.

One immediate consequence of this connection is a straightforward definition of concordance for a risk score containing time dependent covariates. Since the Cox model score statistic is well defined for time dependent terms this justifies calculation of the values C, D, etc in the same way: at each event time the current risk score of the subject who failed is compared to the current scores of all those still at risk.

4 Variance

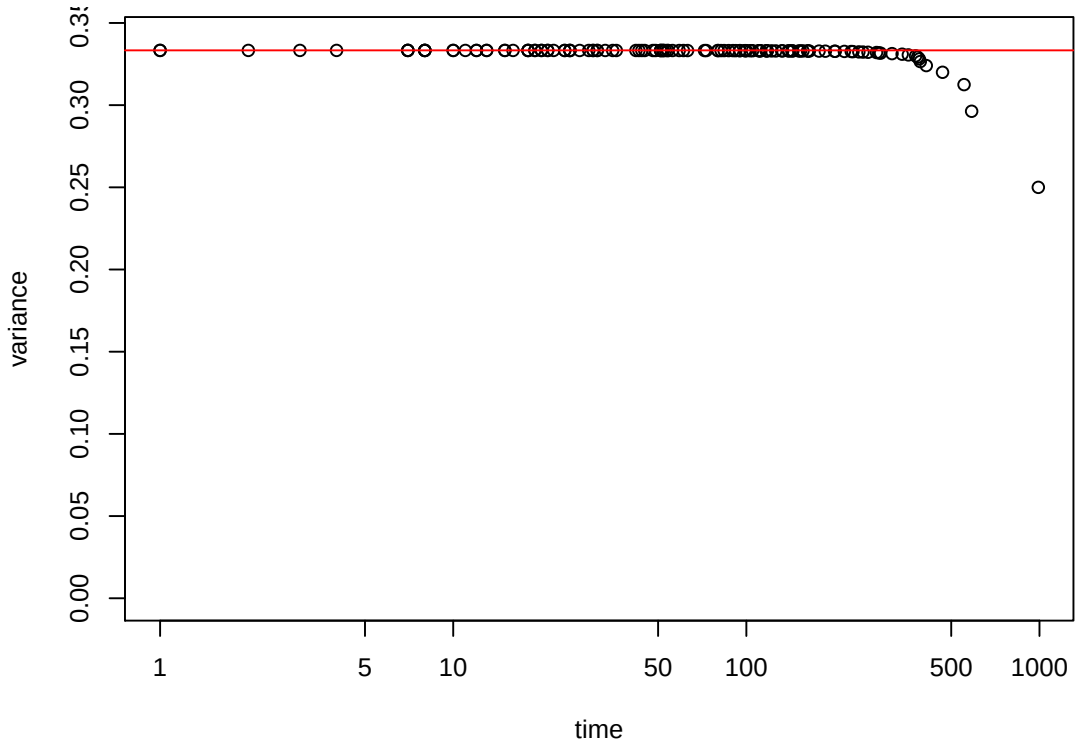
The variance of the statistic is estimated in two ways. The first is to use the variance of the equivalent Cox model score statistic. As pointed out by Watson, this estimate is both correct and efficient under $H_0 : d = 0$, a null concordance of 1/2, and so it forms a valid test of H_0. However, when the concordance is larger than 1/2 the estimate systematically overestimates the true variance. An alternative that remains unbiased is the infinitesimal jackknife (IJ) variance

$$V = \sum_{i=1}^{n} w_i U_i^2$$
$$U_i = \frac{\partial c}{\partial w_i}$$

The concordance routine calculates an influence matrix U with one row per subject and columns that contain the derivitives of the 5 individual counts: concordant, discordant, tied on x, tied on y, and tied on xy pairs. From this it is straightforward to derive the influence of each subject on the concordance, or on any other of the other possible association measures (1) – (4) mentioned earlier. The IJ variance is printed by default but the PH variance is also returned; the earlier `survConcordance` function only computed the PH variance.

The PH variance is a simple sum of the variances for each of the r_i terms shown in the prior plot of the ranks. Under H_0 each of these terms will be approximately uniformly distributed between -1 and 1, which implies that the variance of the individual terms will be about 1/3, and indeed this is true as seen in the plot below. The ranks r are discrete rather than continuous, but this does not have any appreciable affect until the number at risk drops below about 5 or there are a substantial number of ties in the predictor.

```
> plot(variance ~ time, c3$ranks, ylim=c(0, .34), log="x")
> abline(h=1/3, col=2)
```



This suggests that the most statisticially *efficient* estimate of the average rank would use weights of 1 rather than $n(t)$ when adding up the r_i terms, which leads to something very like an ordinary Cox model or logrank test. However, it is no longer the concordance. One aside to remember, however, is the p-value from a Cox model will almost always be smaller than that for the concordance statistic for that same model, which is one side effect of this contrast in efficiency.

5 Multiple concordances

One useful property of using a jackknife variance estimate is that the variance of the difference in concordance between two separately fitted models is also easily obtained. If c_a and c_b are the two concordance statistics and U_{ia} and U_{ib} the influence values, the influence vector for $c_a - c_b$ is $U_a - U_b$. (If subject i increases c_a by .03 and c_b by .01, then he/she raises the difference between them by .02.) It is not even necessary that the models be nested. However, it is crucial that they be computed on the exact same set of observations. Here is a comparison of concordance values from previous models.

```
> ctest <- concordance(fit4, fit5, fit6)
> ctest
Call:
concordance.coxph(object = fit4, fit5, fit6)
```

```
n= 0
     concordance     se
fit4      0.7119 0.0224
fit5      0.7384 0.0210
fit6      0.7359 0.0212

     discordant concordant tied.x tied.y tied.xy
fit4       6261       2529     14     39       0
fit5       6499       2301      4     39       0
fit6       6478       2324      2     39       0
> # compare concordance values of fit4 and fit5
> contr <- c(-1, 1, 0)
> dtest <- contr %*% coef(ctest)
> dvar  <- contr %*% vcov(ctest) %*% contr
> c(contrast=dtest, sd=sqrt(dvar), z=dtest/sqrt(dvar))
  contrast         sd          z
0.02646524 0.01662275 1.59211003
```

5.1 Missing data

The lung data set has several variables with missing data and shows that care is necessary.

```
> colSums(is.na(lung)) # count missing values/variable
     inst      time    status       age       sex   ph.ecog
        1         0         0         0         0         1
 ph.karno pat.karno  meal.cal   wt.loss
        1         3        47        14
> # First attempt
> fit6 <- coxph(Surv(time, status) ~ age + ph.ecog, data=lung)
> fit7 <- coxph(Surv(time, status) ~ meal.cal + pat.karno, data=lung)
> #tryCatch(concordance(fit6,fit7))   # produces an error
>
> # Second attempt
> lung2 <- na.omit(subset(lung, select= -c(inst, wt.loss)))
> fit6b <- coxph(Surv(time, status) ~ age + ph.ecog, data=lung2)
> fit7b <- coxph(Surv(time, status) ~ meal.cal + pat.karno, data=lung2)
> concordance(fit6b,fit7b)
Call:
concordance.coxph(object = fit6b, fit7b)

n= 0
      concordance     se
fit6b      0.6096 0.0284
fit7b      0.5958 0.0286
```

```
      discordant concordant tied.x tied.y tied.xy
fit6b       7435       4733    155     15       0
fit7b       7296       4935     92     15       0
```

6 Weighted concordance

An interesting consequence of the equivalence between the concordance and the Cox model is the question of alternate weightings of the risk scores. Let $0 < r_i(t) < 1$ be the time dependent rank $2(r_i(t) - 1/2)$; the values range from -1 to 1 and their weighted sum is both the Cox score statistic and the numerator of Somers' d. If the original Cox model has a single 0/1 treatment covariate then d is exactly the Gehan-Wilcoxon statistic; replacing these with weights of 1 instead of $n(t)$ will yield the log-rank statistic.

There has been a long debate about the "best" weight for survival tests, and we can apply some of the same historical arguments to the concordance as well. We will point out four of interest:

- Peto and Peto [3] point out that $n(t) \approx n(0)S(t-)G(t-)$, where S is the survival distribution and G the censoring distribution. They argue that $S(t)$ would be a better weight since G may have features that are irrelevant to the question being tested. For a particular data set Prentice [4] later showed that these concerns were indeed justified, and almost all software now uses the Peto-Wilcoxon variant.

- Schemper et al [5] argue for a weight of $S(t)/G(t)$ in the Cox model. When proportional hazards does not hold the coefficient from the Cox model is an "average" hazard ratio, and they show that using S/G leads to a value that remains interpretable in terms of an underlying population model. The same argument may apply more strongly for the concordance, since the target is an "assumption free" assessment of association.

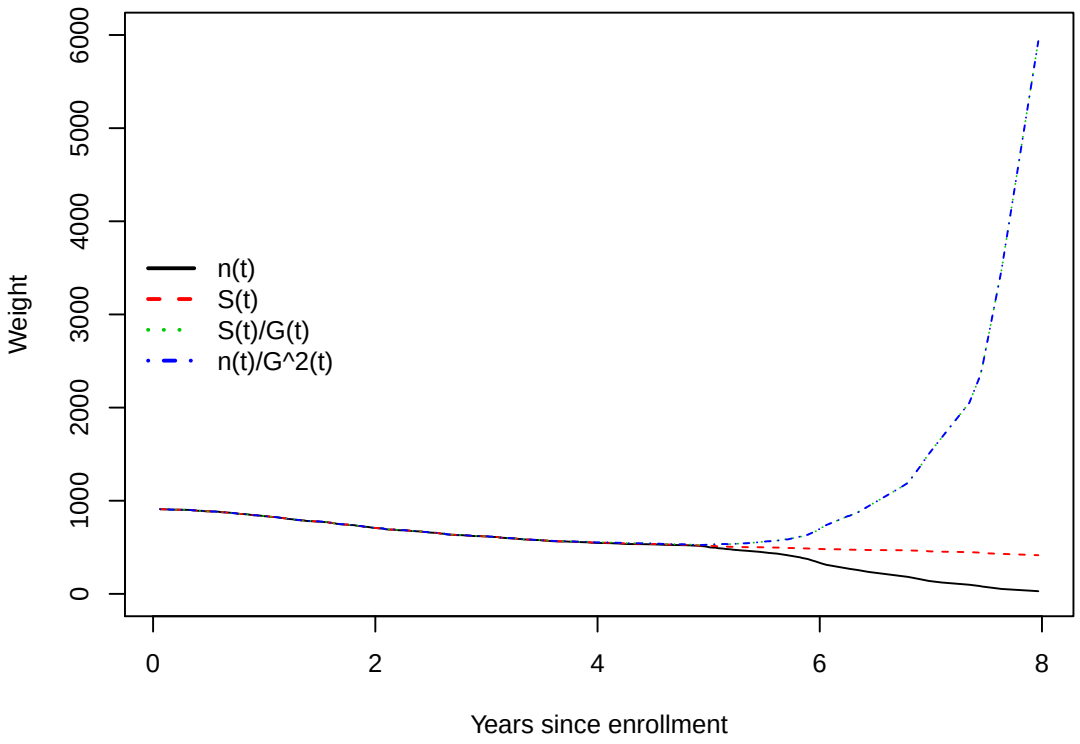
- Uno et al [7] recommend the use of n/G^2 as a weight based on a consistency argument. If we assume that the concordance value that would be obtained after full followup of all subjects (no censoring) is the "right" one, and proportional hazards does not hold, then the standard concordance will not consistently estimate this target quantity when there is censoring. It is "biased". (The Peto and Peto argument might suggest S/G as an equivalent but more stable choice of weight.)

- Korn and Simon [2] point out the importance of restricting the range of comparison. Though a risk model can be used for long-range prediction, in actual patient practice this will not be the case; the model should be evaluated over the time range in which it will actually be used. For example, a physician is quite unlikely to look up my lab tests from 5 years ago, and then compute a 6 year survival probability forward from that point, in order to predict my outcome one year from today.

The `timewt` option allows you to modify weights for the concordance. The options are: `n, S, S/G, n/G, n/G2, I`, the last giving equal weight to each event time. The figure below shows the first four weights for the colon cancer data set. This data is from a clinical trial with 3 years of enrollment followed by 5 years of follow. Since there is almost no one lost to follow-up in the

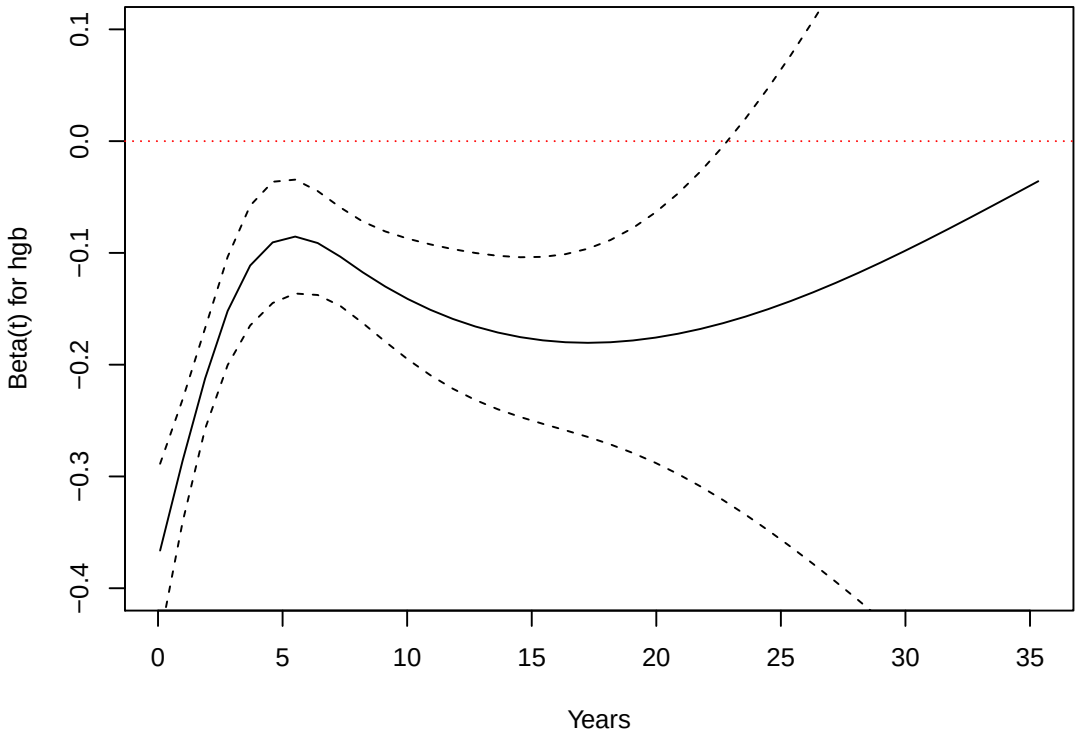
first 5 years all four weights are nearly identical over that time. From 5 to 8 years $S(t)$ continues its steady decline, $n(t)$ plummets due to administrative censoring, and S/G explodes. Even with these changes in weights, the concordance values are all very similar.

```
> colonfit <- coxph(Surv(time, status) ~ rx + nodes + extent, data=colon,
                    subset=(etype==2))   # death only
> cord1 <- concordance(colonfit, timewt="n",    ranks=TRUE)
> cord2 <- concordance(colonfit, timewt="S",    ranks=TRUE)
> cord3 <- concordance(colonfit, timewt="S/G",  ranks=TRUE)
> cord4 <- concordance(colonfit, timewt="n/G2", ranks=TRUE)
> c(n= coef(cord1), S=coef(cord2), "S/G"= coef(cord3), "n/G2"= coef(cord4))
        n         S       S/G      n/G2
0.6555881 0.6543661 0.6535670 0.6535661
> matplot(cord1$ranks$time/365.25, cbind(cord1$ranks$timewt,
                                         cord2$ranks$timewt,
                                         cord3$ranks$timewt,
                                         cord4$ranks$timewt),
          type= "l", ylim= c(0, 6000),
          xlab="Years since enrollment", ylab="Weight")
> legend("left", c("n(t)", "S(t)", "S(t)/G(t)", "n(t)/G^2(t)"), lwd=2,
         col=1:4, lty=1:4, bty="n")
```



A second example where we would expect weights to have a larger influence uses the `mgus2` data set. The underlying study has an unusually long median follow-up time of over 15 years, giving ample time for non-proportionality to manifest. Both creatinine and hemoglobin levels are associated with higher mortality, but of course a 10 year old marker is not nearly as predictive. The R data set has time in months, which has been converted to years for plotting purposes.

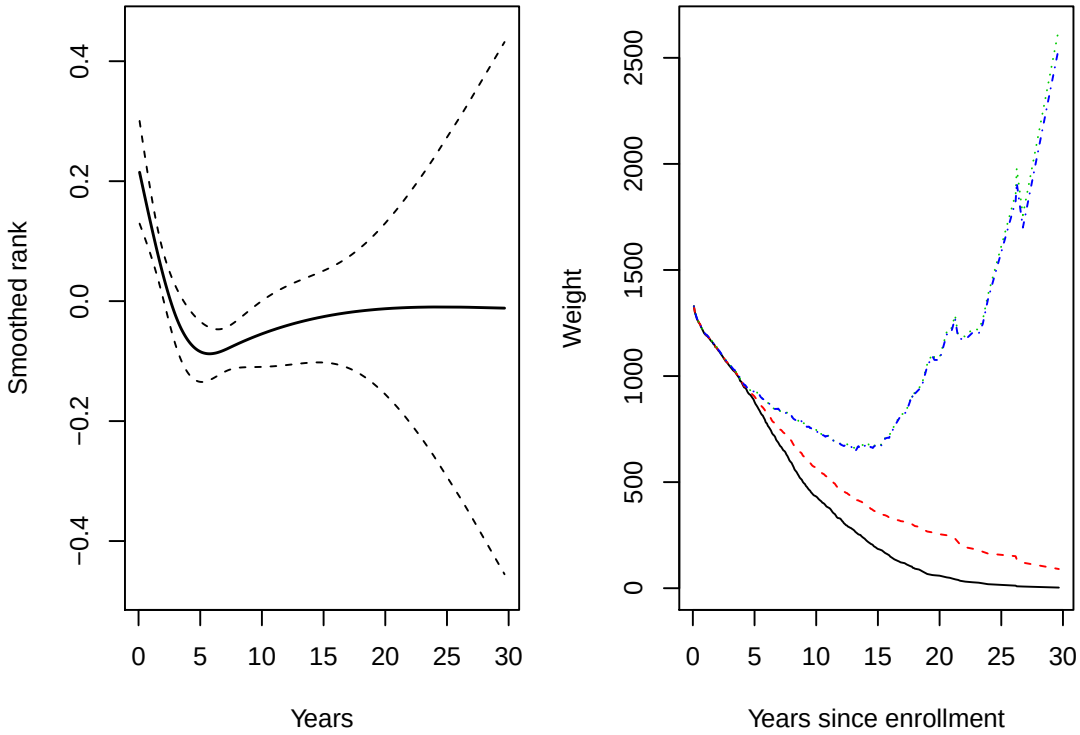
```
> fit6 <- coxph(Surv(futime/12, death) ~ hgb, data=mgus2)
> zp <- cox.zph(fit6, transform="identity")
> plot(zp, df=4, resid=FALSE, ylim=c(-.4, .1), xlab="Years")
> abline(0,0, lty=3, col=2)
```



The predictive effect of hemoglobin drops to less than half after 5 years, with a maximum study follow-up of 35 years. Now calculate the weighted concordance using four approaches.

```
> c6a <- concordance(fit6, timewt="n",    ranks=TRUE)
> c6b <- concordance(fit6, timewt="S",    ranks=TRUE)
> c6c <- concordance(fit6, timewt="S/G",  ranks=TRUE)
> c6d <- concordance(fit6, timewt="n/G2", ranks=TRUE)
> c(n= coef(c6a), S=coef(c6b), "S/G"= coef(c6c), "n/G2"= coef(c6d))
        n         S       S/G      n/G2
0.6065898 0.6038548 0.6011317 0.6011153
> par(mfrow=c(1,2))
> rfit <- lm(rank ~ ns(time,3), data=c6a$ranks)
```

```
> termplot(rfit, se=TRUE, col.se=1, col.term=1,
           xlab="Years", ylab="Smoothed rank")
> matplot(c6a$ranks$time, cbind(c6a$ranks$timewt,
                                c6b$ranks$timewt,
                                c6c$ranks$timewt,
                                c6d$ranks$timewt),
          type= "l",
          xlab="Years since enrollment", ylab="Weight")
> # distribution of death times
> quantile(c6a$ranks$time)
        0%         25%         50%         75%        100%
0.08333333  2.00000000  5.25000000  9.00000000 29.66666667
```



Surprisingly, the four weightings still yield almost identical concordance values. A clue to this lies in the quantile result. The concordance is a weighted mean of the ranks, with one term per death time in the sum. The quantile shows that 1/2 the deaths occur before 5 years and 3/4 of them before 9; the larger weights simply play only a small role in the overll sum.

So, which weight should we use? As shown in the examples above, it may not matter that much. The fact that we have not found examples where the effect is large does not mean there are no such data sets, however. For survival data, one important issue that has not been sorted out is how to extend the weighting arguments to data sets that are subject to delayed entry, e.g., when using age as the time scale instead of time since enrollment. Since this usage is moderately frequent, and also because it is not possible for the `coxph` routine to reliably tell the difference

between such left censoring and simple time-dependent covariates or strata, the default action of the routine is to use the safe choice of $n(t)$ and not to impose a range restriction. Further issues to consider:

1. Consider setting a time (y) restriction using the `ymin` or `ymax` options, based on careful thought about the proper range of interest. This part of the choice may often have the largest practical effect.

2. Safety. If using the usual Gehan-Wilcoxon weights of $n(t)$, the Peto-Wilcoxon variant $S(t)$ would appear advantageous, particularly if there is differential censoring for some subjects.

3. Current data versus future invariance. On the one hand assessment of a model should make use of the all available data, "make the most with what you have", but on the other hand we would like an estimated concordance to stay stable as a study's follow-up matures, which argues for S/G or n/G^2 weights. If $G(t)$ approaches 0, however, these weights can become unstable and so may need to be combined with a time restriction.

4. Equality vs. efficiency. On one hand we would like to treat each data pair equally, but in our quest for ever sharper p-values we want to be efficient. The first argues for $n(t)$ as the weight and the second for using equal weights, since the variances of each term are the same. This is exactly the argument between the Gehan-Wilcoxon and the log-rank tests.

5. For uncensored data n, S and S/G weights are all identical.

Our current opinion is that since the point of the concordance is to evaluate the model in a more non-parametric way, so a log-rank type of focus on ideal p-values is misplaced. This suggests using either S or S/G as weights. Both give more prominence to the later time points as compared to the default $n(t)$ choice, but if time limits have been thought through carefully the difference will often be minor. We most definitely do not agree that the "proper" target for estimation is always the c statistic one would obtain with infinite follow-up and no censoring, which is the unstated assumption underlying Uno's assertion that the ordinary concordance is biased and must be repaired by use of n/G^2. Proportional hazards never is true over the long term, simply because it is almost impossible to predict events that are a decade or more away and thus the points in the $r(t)$ plot above will eventually tend to 0. The starting point should always be to think through exactly *what* one wants to estimate. As stated by Yogi Berra "if you don't know where you are going, you might not get there."

7 Details

This section documents a few details - most readers can skip it.

The usual convention for survival data is to assume that censored values come after deaths, even if they are recorded on the same day. This corresponds to the common case that a subject who is censored on day 200, say, was actually seen on that day. That is, their survival is strictly greater than 200. As a consequence, censoring weights G actually use $G(t-)$ in the code: if 10 subjects are censored at day 100, and these are the first censorings in the study, then an event on day 100 should not be given a larger weight. (Both the Uno and Schemper papers ignore this detail.)

When using weights of $S(t)$ the program actually uses a weight of $nS(t)$ where n is the number of subjects. The reason is that for a stratified model the weighted number of concordant, discordant and tied pairs is calculated separately for each stratum, and then added together. If one stratum were much smaller or larger than the others we want to preserved this fact in the sum.

Consider a time point t at which there were 3 events out of 40 subjects at risk. For the the ordinary concordance the time weight at this point will be the number of *comparable pairs* for each of the 3 events, i.e., 37 for each. The rank of an event will be 1 if its predictor is smaller than all 37 comparitors and -1 if it is larger than all 37. When using a weight of $S(t)$ the weight will be $nS(t)$ at that point, for S/G it will be $nS(t)/G(t-)$. The Cox model computation takes a slightly different view, in that all 40 subjects are considered to be at risk for each of the 3 events. The upshot of this is that the time weight changes to 40 rather than 37, while the rank becomes smaller by a factor of 37/40. The weighted sum remains the same, i.e., the Cox score statistic is equal to the numerator of the concordance statistic. For other weights $S(t-)$ replaces $S(t)$ in the Cox calculation, the individual ranks again shrink slightly but the weighted sum stays the same.

References

[1] Frank E Harrell, Kerry L Lee, and Daniel B Mark. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat. in Medicine*, 15:361–387, 1996.

[2] E. L. Korn and R. Simon. Measures of explained variation for survival data. *Stat. in Medicine*, 9:487–503, 1990.

[3] R. Peto and J. Peto. Asymptotically efficient rank invariant test procedures (with discussion). *J. Royal Stat. Soc. A*, 135(2):185–206, 1972.

[4] Ross L Prentice and P Marek. A qualitative discrepancy between censored data rank tests. *Biometrics*, 35(4):861–867, 1979.

[5] M. Schemper, S. Wakounig, and G. Heinze. The estimation of average hazard ratios by weighted Cox regression. *Stat. in Medicine*, 28(19):2473–2489, 2009.

[6] T. M. Therneau and D. A. Watson. The concordance statistic and the Cox model. Technical Report 85, Department of Health Science Research, Mayo Clinic, 2015.

[7] H. Uno, T. Cai, M. J. Pencina, R. B D'Agnostino, and L. J. Wei. On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Stat. in Medicine*, 30(10):1105–1117, 2011.