Appendix A

Formulas

This section gathers all the notation and formulas for the models in one place. For the data we use counting process notation:

$N_i(t)$	cumulative number of events for subject i, up to time t
$N_{ijk}(t)$	in a multistate model, the number of transitions from state j to state k for subject i
$Y_i(t)$	0/1 indicator that subject i is at risk at time t
$Y_{ij}(t)$	in a multistate model, the indicator that subject i is at risk for a transition out of state j
X	n by p matrix of predictors where p is the number of predictors, X_i the row vector for observati
w	vector of case weights for the subjects, W a diagonal matrix of the weights
β	vector of estimates
$\eta = X\beta$	vector of linear predictors

A.1 Nelson-Aalen and Kaplan-Meier estimates

The most common estimates of the cumulative hazard and survival are the Nelson-Aalen and Kaplan-Meier estimates.

$$\begin{aligned}
\hat{\lambda}(t) &= \frac{\sum w_i dN_i(s)}{\sum_i w_i Y_i(s)} \\
\hat{\Lambda}(t) &= \int_0^t \hat{\lambda}(s) \\
\text{var}(\hat{\Lambda}(t)) &= \int_0^t \frac{\sum w_i dN_i(s)}{\left(\sum_i w_i Y_i(s)\right)^2} \qquad \text{(A.1)} \\
\widehat{S}(t) &= \prod_{s \le t} \left(1 - \frac{\sum w_i dN_i(t)}{\sum_i Y_i(t)}\right) \\
&= \prod_{s \le t} \left(1 - \hat{\lambda}(s)\right) \qquad \text{(A.2)} \\
\text{var}(\widehat{S}(t)) &= \widehat{S}^2(t) \int_0^t \frac{\sum w_i dN_i(s)}{\sum_i w_i Y_i(s) \left[\sum_i w_i (Y_i(s) - dN_i(s))\right]} \qquad \text{(A.3)} \\
M_i(t) &= N_i(t) - Y_i(t)\hat{\Lambda}(t)
\end{aligned}$$

The variance estimate (A.3) for the Kaplan-Meier is known as the Greenwood estimator. An alternate variance is $\widehat{S}^2(t)\text{var}(\hat{\Lambda}(t))$, but it has been found to be inferior to the Greenwood.

The quantity $M_i(t)$ is the martingale residual process for subject i, it is a running total of the observed - expected number of events for the subject. (Since the theory of martingales was motivated by cumulative winnings in games of chance, it's applicability to cumulative events is perhaps not surprising). At any time point $\sum w_i M_i(t) = 0$, and M_i without a time indicator is understood to be the terminal value.

When there are multiple events tied at a single time, an alternate estimate of the cumulative hazard can be based on the idea of *coarsened* data. That is, if the time scale had been more finely measured the ties would not have occurred. Say that there were 10 subjects at risk of which 3 had an event. Using the data, observed on a coarsened time scale, the Nelson-Aalen estimate will have a jump of 3/10, but if time had been measured continuously the jump would have been 1/10 + 1/9 + 1/8 = .336. This estimate is explored by Fleming and Harrington [36].

In the more general setting of case weights, if there were d tied events at time t we can apply the Nelson-Aalen estimate to a synthetic dataset in which there are d distinct event times t, $t-\epsilon$, $t-2\epsilon$, ..., and each of the tied observations is split into d synthetic subjects with weight w_i/d, one of which has an event at the first of the synthetic times, one at the second synthetic time, etc. The idea is that each subject could have been the first, second, third, etc of the d events, so we spread each of them out evenly over those possibilities. Computationally, we do not need to actually create the synthetic data. If there are d events at some time t, let $n_2(t) = \sum w_i dN_i(t)$ be the sum of weights for those who have an event at that time and $n_1(t) = (\sum w_i Y_i(t)) - n_2()$ be the sum for all the others. The the increment to the hazard estimate and the derivative are

$$d\hat{\Lambda}(t) = \frac{n_2(t)}{d} \sum_{m=1}^{d} 1/(n_1 + (m/d)n_2) \tag{A.4}$$

$$\frac{\partial \hat{\Lambda}(t)}{\partial w_i} = \int_0^t dN_i(s) d\hat{\Lambda}(s) - Y_i(s) \sum_{m=1}^{d} \frac{(m/d)I(i \in n_2) + 1I(i \in n_1)}{(n_1 + (m/d)n_2)^2} \tag{A.5}$$

$$\neq \int_0^t \frac{dM_i(s)}{\sum_i w_i Y_i(s)} \tag{A.6}$$

From equation (A.4), by moving the divisor d to the right we see that the method is essentially using an average denominator for the FH increment. The step from (A.5) to (A.6) that was used in the Nelson-Aalen case fails here since products do not factor — $\sum a_i b_i \neq (\sum a_i)(\sum b_i)$; one can't separately sum the dM contributions at a tied event time.

The coarsening argument does not affect the Kaplan-Meier estimate, which will be 7/10 on the coarsened scale and the product (9/10)(8/9)(7/8)= 7/10

for continuous time, i.e., the solution does not change. This is true for weighted data as well.

For a continuous survival distribution $S(t) = \exp(-\Lambda(t))$. Since $\exp(-x) \approx 1 - x$ when x is small, this is also approximately true for the non-parametric estimates: $\widehat{S}(t) \approx \exp(-\hat{\Lambda}(t)]$. In this case the exponent of the Fleming-Harrington hazard estimate will be closer to the KM than the exponent of the Nelson-Aalen [36], but the differences between all three of KM, exp(-NA) and exp(-FH) are normally very small until the number at risk becomes small, at which point the standard errors of the curves are large as well. The exponential estimate plays an important role in curves for a Cox model, but is rarely used for simple survival.

A.2 Aalen-Johansen estimate

The multistate analog to the Kaplan-Meier curve is the Aalen-Johansen estimator. It estimates $p(t)$, a vector containing the probability of being in each of the states at time t. Mathematically the estimate is simple. For each unique time that an event occurs form a transition (or hazards) matrix $H(t)$ with elements $\hat{\lambda}_{jk}(t) =$ the fraction of subjects who transition from state j to k at time t, among those in state j just prior to t. Formally

$$\hat{\lambda}_{jk}(t) = \frac{\sum_i w_i Y_{ij}(t) dN_{ijk}(t)}{\sum_i w_i Y_{ij}(t)}, \; j \neq k$$

where $Y_{ij}(t)$ is 1 if observation i is in state j at time $t-$ (just before the event), $dN_{ijk}(t)$ is 1 for a transition from state j to k, and w is an optional case weight. The diagonal elements of H are such that the row sums of H are 1.

Then

$$p(t) = p(0) \prod_{s \leq t} H(s) \tag{A.7}$$

where $p(0)$ is the initial distribution of subjects. H is equal to the identity matrix at any time point without an observed transition, so we only need to include transition times in the product.

The elements of $p(t)$ sum to 1 at each time point: everyone has to be somewhere. Likewise, the rows of each transition matrix H sum to 1: everyone has to either stay put (diagonal) or transition to a different state. For the simple alive→dead model there are only two states and H has the simple form

$$H(s) = \begin{pmatrix} \frac{\sum Y_i(s) - dN_i(s)}{\sum_i Y_i(s)} & \frac{\sum dN_i(s)}{\sum_i Y_i(s)} \\ 0 & 1 \end{pmatrix}$$

$H_{21} = 0$ since no one transitions from dead to alive; death is an *absorbing* state. Writing out the matrices for the first few transitions and multiplying

them leads to

$$p_1(t) = \prod_{s \le t} \frac{\sum Y_i(s) - \sum dN_i(s)}{\sum Y_i(s)} \tag{A.8}$$

which we recognize as the Kaplan-Meier estimate of survival.

An alternate estimator is based on matrix exponentials

$$\begin{aligned} p(t) &= p_0(t) \prod_{s \le t} \exp(H(s) - I) \\ &= p_0(t) \prod_{s \le t} \exp(A(s)) \end{aligned} \tag{A.9}$$

The off diagonal elements of A are $\hat{\lambda}_{jk}(t)$, the same as H, but the diagonal element are constrained so that rows sum to zero. We normally cannot further collapse equation (A.9), since in general $\exp(C)\exp(D) \neq \exp(C + D)$ for two matrices C and D. We refer to (A.7) and (A.9) as *direct* and *exponential* estimates of the probability in state matrix. For single transition models these correspond to the Kaplan-Meier and Fleming-Harrington estimates, respectively. The $A(s)$ notation is common in the counting process literature, in which case $A(s) + I$ usually replaces $H(s)$.

A.3 Cox Model

Let $r_i(t) = \exp(\eta_i(t) - c) = \exp(X_i(t)\beta - c)$ be the risk score for each subject at time t, with c an arbitrary centering constant. Mathematically we can set $c = 0$, since it cancels out of the partial log-likelihood (LPL), but computationally it is important to choose a value that avoids extreme arguments to the exponential function. (The value c is universally left out of textbook formulas, and at the same time universally recognized as crucial by writers of code.) A common choice is the mean risk score $(1/n)\sum \eta_i$, $\eta = X\hat{\beta}$, but the exact centering value is not critical.

We have the following quantities:

$$LPL = \sum_i \int \left(w_i \log(r_i(s)) - \log \left[\sum_j Y_j(t) w_j r_j(s) \right] \right) dN_i(s) \tag{A.10}$$

$$\hat{\lambda}(t; c) = \frac{\sum w_i dN_i(t)}{\sum_i Y_j(t) w_i r_i(t)} \tag{A.11}$$

$$\hat{\Lambda}(t; c) = \int_0^t \hat{\lambda}(s) \tag{A.12}$$

$$M_i(t) = N_i(t) - Y_i(t) r_i \hat{\Lambda}(t)$$

$$\overline{x}(t) = \frac{\sum_i Y_i(t) w_i r_i(t) X_i(t)}{\sum_i Y_j(j) w_i r_i(t)}$$

$$U = \sum_i \int (x_i - \overline{x}(s)) dN_i(s) \tag{A.13}$$

$$H = \sum_i \int dN_i(t) \frac{\sum_{j=1}^n w_j r_j(s) (X_j(s) - \overline{x}(s))^2}{\sum_j Y_j(s) w_j r_j(s)} dN_i(s) \tag{A.14}$$

$$\tag{A.15}$$

Since $N(t)$ is a step function, integrals with respect to dN are equivalent to a sum that contains a term at each death time. Each term of the log partial-likelihood (LPL) compares the risk score r_i of a subject who had an event to the sum over all those at risk. The baseline hazard estimate (A.12) is for a hypothetical subject with risk score $r = 0$ or equivalently $X\hat{\beta} = c$. Textbooks commonly use the formula with $c = 0$ and refer to the result as *the* baseline hazard $\hat{\lambda}_0$, but again, this is a very unwise choice in computer code due to potential numeric errors.

When $\beta = 0$ then $\hat{\Lambda}$ is equal to a Nelson-Aalen estimate, but in this context is more often referred to as the Breslow estimate. $\overline{x}$ and H are the weighted mean and variance of the covariate vectors at each event time. The total of $H(t)$ over the death times is the second derivative of the LPL, also known as the Hessian or information matrix. U is the contribution to the first derivative of the LPL.

The martingale residual is the observed - expected number of events for each subject i. M_i without an argument is taken to be the value at the maximum follow-up for the subject.

Three important identities are

$$\sum_i w_i M_i(t) = 0$$

$$M_i = \frac{\partial LPL}{\partial \eta_i} \tag{A.16}$$

$$\begin{aligned} U &= \frac{\partial LPL}{\partial \beta} \\ &= M'WX \end{aligned} \tag{A.17}$$

where W is the diagonal matrix of weights. Equation (A.17) plays an important role in MCMC or machine learning approaches which make use of first derivative information, and for which X may be both large and sparse. Since the marginale residual M can be computed in $O(n)$ steps this allows for fast sparse matrix calculation of the derivative.

A.3.1 Tied event times

The theory underlying the Cox model is derived for the case of continuous t, but in real data there often are tied event times. There are four primary approaches to handle ties. The simplest is simply to ignore them and continue to use the formulas just above. The upshot is that if there are d events at some time t, each of these d subjects is considered to be at risk for all d of the events. This is known as the Breslow estimate of $\hat{\beta}$; the corresponding estimate of cumulative hazard $\hat{\Lambda}$ is called the Breslow estimate of the baseline hazard.

When dealing with a terminal event such as death the above calculation is flawed by the fact that subjects must by definition leave the risk set when the event occurs. That is, if we assume the true data is continuous then there is *some* true order in which the deaths occurred; whomever died first cannot have been at risk for the later events. There are four common approaches to resolve the issue, of which the first is to ignore it, i.e., the Breslow estimate. The Efron approximation is based on the same coarsened data logic as the Fleming-Harrington estimate of the cumulative hazard: if there are d tied deaths, imagine d separate times separated by a small increment ϵ; subject 1 has $1/d$ chance of being the first death, $1/d$ of being the second death, ..., and similarly for the other ties. Mechanically, one can create d distinct synthetic times clustered at the tied event time, and divide each of the tied events into d synthetic subjects with weights w_i/d, one portion of the first subject perishes at the first synthetic time, one at the second synthetic time, etc. Consequently all d of the synthetic clones of a subject are at risk at the first event time, $d-1$ of them will be at risk at the second giving a total weight of $w_i(d-1)/d$, ... Computationally, divide the LPL denominator sum into two portions where $s_2(t)$ is the sum over the tied events and $s_1(t)$ the sum over the remaining subjects at risk, then the first and second terms of the increment to the LPL

at time t are

$$\text{first term} = \sum_i w_i dN_i(t)(\eta_i - c)$$

$$\text{second term} = \sum_{k=1}^{d} -\log[n_1(t) + (k/d)n_2(t)]$$

The first term is unchanged from the Breslow approach while the values in the the second term are decreased.

The estimate of the cumulative hazard will also be changed, essentially becoming a Fleming-Harrington increment rather than a Breslow increment at each time. Although Efron [34] did not discuss estimation of the cumulative hazard, that paper being focused on properties of $\hat{\beta}$, for consistency we refer to this as the Efron estimate of the hazard function. Using this approach the martingale identities (A.16) and (A.17) still hold. Though it is more work to create the computer code for the Efron approach than it is for the Breslow approximation, the compute time for the Efron and Breslow approaches is essentially identical.

Some software packages use an Efron approach to compute the likelihood and $\hat{\beta}$, but then switch to the Breslow estimate of the baseline hazard when computing the residual, e.g. SAS. For this hybrid estimate equations (A.16) and (A.17) will no longer hold; this also impacts the robust variance estimate. The hybrid residuals still sum to zero.

A third approach to ties is to compute the exact partial likelihood (EPL), as found in Cox's original paper. This approach views the time scale as discrete. The first term in the LPL is unchanged from before, but if there are d events out of n subjects at risk the second term is now a sum over all $\binom{n}{d}$ ways of choosing a subset of size d. For even modest values of n and d this can be a *very* substantial computation. A clever nested algorithm [41] speeds this up considerably, but it is not clear that the EPL is actually worth the all the effort. Many software packages implement this approximation, but none (that the authors are aware of) go through the extra work of computing a matching hazard estimate, martingale residual, or robust variance. In practice the solution using an Efron approximation is often very close to the EPL estimate.

A fourth approach due to Prentice takes advantage of an algebraic identity, namely that the EPL summation for a given term of the partial likelihood is precisely the analytical solution to a particular integral after d nested applications of an integration by parts formula. The code can then evaluate said integral using numerical integration. This last is implemented in only a few packages.

The authors' opinion at the current time is that

1. When ties are infrequent the approximation for ties has no impact of consequence. Use whichever computation is convenient, e.g., the default for whatever software you favor.

	Few ties		Moderate ties		Heavy ties	
	coef	se	coef	se	coef	se
Breslow	0.48	0.11	0.46	0.11	0.41	0.11
Efron	0.48	0.11	0.48	0.11	0.46	0.11
Exact	0.48	0.11	0.5	0.12	0.53	0.13

Table A.1 *Approximations for ties using the lung dataset from the survival package, with ph.ecog as the single covariate. Few ties: original data; there are 115 unique death times, 22 with 2 deaths, and 2 with 3 tied deaths. Moderate ties: replace time with floor(time/30); there are 1–15 deaths at each unique death time. Heavy ties: replace time with floor(time/100); there are only 10 unique death times.*

2. When there are a moderate number of ties or if one wants to be more "pure" about the underlying assumption of continuous time, then use a formal correction. Even then, the numeric difference will often be slight, i.e., less than one fifth the standard error of the estimate. Table A.1 gives a concrete example.
3. When time is actually discreet and there are a large number of ties a good argument can be made for the exact partial likelihood (EPL). However, the calculation quickly becomes intractable, and the Efron approximation is often sufficiently close in value.
4. Once an estimate is chosen, be consistent. An argument that a hybrid algorithm which mixes an Efron or EPL partial likelihood along with a Breslow hazard is "numerically close enough" immediately begs the question of why one would not use the simpler Breslow approach throughout.
5. The entire discussion becomes much more complex for a multistate model, and it is no longer as clear what is accomplished by each of the approximations. In that case simplicity may be the best guide, i.e., use the Breslow estimate.
6. Our overall advice is not to worry about ties. Though at one time this topic generated substantial thought and interest, it simply isn't very important.

A.3.2 Absolute risk

Predicted survival curves $S(t)$ from a Cox model, or equivalently probability in state estimates $p(t)$ from a multistate model, are most often based on an analog of the exponential estimate (A.19) rather than the direct estimate (A.18). Predicted curves are for a hypothetical subject with chosen covariates z. Let $A(s; z)$ have off diagonal elements $\hat{\lambda}(s; z)$, and then fill in the diagonal element such that row sums of A are zero. That is, $\hat{\lambda}$ with centering constant

$c = z\beta$. The direct and exponential estimates of the probability in state are

$$\hat{\lambda}(t;z) = \frac{\sum_i dN_i(t)}{\sum_i Y_i(t)e^{(X_i - z)\beta}}$$

$$\hat{p}(t;z) = p(0)\prod s \le t(A(s;z) + I) \tag{A.18}$$

$$\hat{p}(t;z) = p(0)\prod s \le te^{A(s;z)} \tag{A.19}$$

Unless the number of subjects at risk is very small (< 10) the two estimates will often be equivalent for practical purposes. One issue with the direct estimate occurs for predictions of high risk subjects: the diagonal of $A + I$ may be negative, which is an invalid transition matrix (more than 100% of the subjects are predicted to make a transition). The exponential form avoids this error. The issue normally occurs only when the number at risk is very small (often ≤ 5). Most often this at the far right tail of the curve, a point where the standard deviation of the estimate is large. Because of the large se a precise answer for further time points may not matter and some software will truncate the estimate at this point as a way to go forward.

For a simple alive-dead model the equations simplify to

$$\hat{\Lambda}(t;z) = \int_0^t \hat{\lambda}(s;z)ds \tag{A.20}$$

$$= e^{z\beta}\int_0^t \frac{\sum_i dN_i(s)}{\sum_i Y_i(s)e^{X_i\beta}} \tag{A.21}$$

$$S(t;z) = \exp(-\hat{\Lambda}(t;z)) \tag{A.22}$$

$$S(t;z) = \prod_{s \le z}\left(1 - \hat{\lambda}(s;z)\right) \tag{A.23}$$

Equation (A.21) often appears in print, with the right hand term (for $z = 0$) labeled as the "baseline hazard", but in computation it is wiser to use (A.20). The exponential estimate (A.22) is known as the Breslow estimate, and is uniformly used. The direct estimate A.23 can generate a negative survival estimate for z values corresponding to a high risk subject ($\hat{\lambda}(t;z) > 1$). An alternate estimate due to Kalbfleisch and Prentice avoids this issue but is only applicable to the simple survival case. It has the form

$$\hat{S}(t;z) = \prod_{s \le t}\alpha_s$$

$$\sum \frac{dN_i(s)r_i}{1 - \alpha_{s_i}^{r}} = \sum Y_i(s)r_i$$

$$r_i = e^{(X_i - z)\beta}$$

where the individual terms $\alpha(s)$ satisfy the second equation. If $\beta = 0$ it agrees with the Kaplan-Meier.

The fact that the exponential form is the most prominent for single state Cox models is at first surprising, since the exponential form is almost never used for non-parametric curves. This is, we think, as much due to the fact that it was easier to implement than any statistical argument. For multistate curves, we will see below that the variance is somewhat harder to compute for the exponential form, which has led to a higher prevalence of the direct estimate in initial software.

The standard variance for the cumulative hazard for a subject with covariate vector z has two terms $A + B$. The first is an analog of the Nelson-Aalen variance and the second accounts for the variance in $\hat{\beta}$.

$$\begin{aligned} A(t) &= r_z^2 \int_0^t \frac{\sum w_i dN_i(s)}{\left(\sum_i w_i r_i(s) Y_i(s)\right)^2} \\ B(t) &= d(t)' \mathcal{I}^{-1} d(t) \\ d(t) &= \frac{\partial \hat{\Lambda}_z(t)}{\partial \beta} \\ &= r_z \int_0^t (\overline{x}(s) - z) d\hat{\Lambda}(s) \end{aligned}$$

The integral in the last line can be recognized as -1 times the score residual process for an observation with covariate z, which measures the effect of each subject on the first derivative of $\hat{\beta}$. The risk score $r_z = \exp(z'\beta - c)$ is centered, while $\overline{x}(s) - z$ is independent of centering.

A.3.3 Matrix exponential and multistate models

Absolute risk estimates require the matrix exponential of the transition matrix to be computed, sometimes hundreds of times. For matrices A and B, $\exp(A)\exp(B) \neq \exp(A+B)$, so formula (A.19) does not have a shortcut, i.e., you cannot use the matrix exponential of the cumulative hazard. The matrix exponential is formally defined as

$$\exp(A) = I + \sum_{j=1}^{\infty} A^i / i! \tag{A.24}$$

The computation is nicely solved by the expm package in R *if* we didn't need derivatives and/or high speed. We want both.

We can break this down into 4 main cases.

1. When there is only one event at a particular time point, say from state j to state k, then A contains zero for all but the jj and jk elements, $\exp(A)$ will equal the identity matrix for all but element jj, which will be $\exp(-\hat{\lambda}_{ij}(t))$, and element jk will be 1 minus this. That is, the computation is no different than a simple exponential, and the derivative is likewise simple. For many datasets this case will hold for almost all time points.

2. If there are multiple events at the time point, but all all share the same initial state then the formula, shown below, is likewise simple. This will be the case for competing risk models, for instance.
3. If $A = BDB^{-1}$ where D is diagonal, then $\exp(A) = B\exp(D)B^{-1}$, the exponential of a diagonal matrix is simply a diagonal matrix of the elementwise exponentials. The derivative also has a simple form [57]. If a model is acyclic (no loops) it can be arranged so that A is upper triangular, and the solution found using a generalized cholesky decomposition.
4. In the general case we use a Pade-Laplace algorithm, which is the standard method for the matrix exponential, with additions to compute the derivative.

The fact that A is a rate matrix, i.e., each row sums to zero and the off diagonal elements are non-negative, implies that $\exp(A)$ will be a transition matrix: all elements are non-negative and each row sums to 1. It also means that many of the edge cases which can plague a general matrix exponential algorithm will not arise.

If there is only a single departure state j, then all rows of A except row j will be zero. It is easy to verify that the matrix power A^i also has zeros in all rows but the jth. Equation (A.24) can then be applied element by element to work out the result. In the below assume $k \neq j$.

$$\begin{aligned}
\left(e^A\right)_{jj} &= e^{A_{jj}} \\
\left(e^A\right)_{jk} &= (e^{A_{jj}} - 1)A_{jk}/A_{jj} \\
\left(e^A\right)_{kk} &= 1 \\
\left(e^A\right)_{kl} &= 0
\end{aligned}$$

As a specific case consider a competing risks model with the states labeled as 1–3 and 1 for the initial state. Only the first row of A will be non-zero, and rows 2–3 of both $(A + I)$ and $\exp(A)$ match an identity matrix. In closed form, the estimated fraction in state 2 for the direct estimate is

$$S(t) = \prod_{s \leq t} 1 - [\hat{\lambda}_{12}(s) + \hat{\lambda}_{13}(s)] \tag{A.25}$$

$$p_2(t) = \int_0^t \hat{\lambda}_{01}(s) S(s-) \tag{A.26}$$

$$\tag{A.27}$$

and for the exponential estimate

$$\hat{\Lambda}(t) = \sum_{s \leq t} \hat{\lambda}_{12}(s) + \hat{\lambda}_{13}(s) \tag{A.28}$$

$$p_2(t) = \int_0^t \hat{\lambda}_{01}(s) e^{-\hat{\Lambda}(s-)} \left[\frac{e^{A_{11}} - 1}{A_{11}}\right] \tag{A.29}$$

The standardly quoted formula for the cumulative incidence is, interestingly, not quite either of them.

A.4 Robust variance

The survival package allows for a robust variance for most of the estimates; for multistate models or datasets where a subject can have more than one event the robust variance is essential. Computations are based on the infinitesimal jackknife (IJ) estimate [35]. Let $\hat{\beta}$ be an estimate, which could be the coefficient of a regression model, the value of a survival curve, etc., and let w_i be a case weight or sampling weight for each subject. The leverage matrix D is then defined to have elements

$$D_{ij} = \left. \frac{\partial \hat{\beta}_j}{\partial w_i} \right|_w$$

One interesting property of D is that column sums are 0, which can act as a useful check on computations.

The IJ variance matrix is then

$$V_{IJ} = D'W^2D$$

where W is a diagonal matrix of observation weights; most commonly the weights are 1. The simple grouped jackknife replaces the central W^2 term with $WBB'W$ where B is an n by g 0/1 grouping matrix. B is essentially the design matrix for a linear model which had the grouping variable as a single prediction factor. More complex sampling designs can be realized with other replacements for B [14] but this topic is outside of our scope. (Many of the package routines will return the weighted influence matrix WD, however, so users can wrap their own.)

The IJ estimate is familiar in multiple statistical contexts, under multiple names. In a generalized estimating equations models the simple grouping matrix B leads to the *working independence* estimate of variance. It also arises in survey sampling as the Horvitz-Thompson estimate. For a linear regression model, the infinitesimal jackknife approach leads to the estimate

$$D'D = (X'X)^{-1}X'RX(X'X)^{-1}$$

of White [126, 127], where R is a diagonal matrix containing the squared residuals. White recommends its use when the data are heteroscedastic. If one believed the data to be homoscedastic, then a natural step would be to replace R with $\hat{\sigma}^2 I$, the "average" squared residual times an identity matrix. The estimator then collapses down to the usual linear model variance estimate. We can rewrite $D'D$ in the sandwich form as well,

$$\hat{\sigma}^2(X'X)^{-1}\ [\hat{\sigma}^{-2}X'RX\hat{\sigma}^{-2}]\ \hat{\sigma}^2(X'X)^{-1},$$

a nonparametric variance estimator sandwiched between two copies of the usual variance matrix $\hat{\sigma}^2(X'X)^{-1}$.

A.4.1 Nelson-Aalen and Kaplan-Meier estimates

$$\begin{aligned}
\hat{\Lambda}(t) &= \int_0^t \frac{\sum w_i dN_i(s)}{\sum_i w_i Y_i(s)} \\
\text{var}(\hat{\Lambda}(t)) &= \int_0^t \frac{\sum w_i dN_i(s)}{\left(\sum_i w_i Y_i(s)\right)^2} \qquad \text{(A.30)} \\
\frac{\partial \hat{\Lambda}(t)}{\partial w_j} &= \int_0^t \frac{dN_j(s) - Y_j(s)/\sum_i w_i Y_i(s)}{\sum_i w_i Y_i(s)} \\
&= \int_0^t \frac{dM_j(t)}{\sum_i w_i Y_i(s)} \qquad \text{(A.31)} \\
\widehat{S}(t) &= \prod_{s \le t} \left(1 - \hat{\lambda}(s)\right) \qquad \text{(A.32)} \\
\text{var}(\widehat{S}(t)) &= \widehat{S}^2(t) \int_0^t \frac{\sum w_i dN_i(s)}{\sum_i w_i Y_i(s) \left[\sum_i w_i (Y_i(s) - dN_i(s))\right]} \qquad \text{(A.33)} \\
\frac{\partial \widehat{S}(t)}{\partial w_j} &= \frac{\partial \widehat{S}(t-)}{\partial w_j}[1 - \hat{\lambda}(t)] - \widehat{S}(t)\frac{\partial \hat{\lambda}(t)}{\partial w_j} \qquad \text{(A.34)}
\end{aligned}$$

The usual variance estimates for the Nelson-Aalen estimate of the cumulative hazard and for the Kaplan-Meier are given in (A.30) and (A.33). Robust variance estimates based on (A.31) and (A.34) will be appropriate when there are observation weights or grouping structure. The full set of influence estimates J would have a row for each subject and a column for each event time, and $J'WJ$ would then be an estimate of the full variance-covariance matrix of all the time points. This is impractical and unnecessary as we normally only want variances for each time point separately. For this is suffices to retain a vector of per-observation influence values, at the current time, updating as we move forward in time.

The Fleming-Harrington estimate of hazard and its derivative are below. In this case the first derivative does not admit of the final simplification of equation (A.31) above.

$$\begin{aligned}
\hat{\lambda}(t) &= \frac{n_2(t)}{d} \sum_{m=1}^{d} 1/(n_1 + (m/d)n_2) \\
\frac{\partial \hat{\Lambda}(t)}{\partial w_i} &= \int_0^t dN_i(s) d\hat{\Lambda}(s) - Y_i(s) \sum_{m=1}^{d} \frac{(m/d)I(i \in n_2) + 1I(i \in n_1)}{(n_1 + (m/d)n_2)^2} \\
&\neq \int_0^t \frac{dM_i(s)}{\sum_i w_i Y_i(s)}
\end{aligned}$$

A.4.2 Aalen-Johansen

The multistate analog to the Nelson-Aalen hazard estimator is no different than the the cumulative hazard for a single state: each transition from a state j to a state k is computed separately, independent of the other transitions. The derivatives of the multistate hazard and cumulative hazard thus follow directly. The absolute risk estimate $p(t)$ involves matrix products so is a bit more work.

Assume s states and let $p(t)$ be a vector of length s with $p_j(t) =$ probability that the process is in state j at time t. Then for the direct estimate

$$p(t) = p(0) \prod_{s \le t} (A(s) + I)$$

$$\frac{\partial p(t)}{\partial w_i} = U(t) \tag{A.35}$$

$$= U(t-)(A(t) + I) + p(t-)\frac{\partial A(t)}{\partial w_i} \tag{A.36}$$

U is a matrix with one row per observation and s columns.

Let $n_j(t) = \sum_i Y_{ij}(t) w_i$ be the sum of weights for subjects in state j at time $t-$. In the transition matrix $A(t)$ each observation at risk with initial state j and final state k at time t adds $w_i/n_j(t)$ to the jk element of A. Thus

$$\frac{\partial A_{jk}(t)}{\partial w_i} = Y_{ij}(t) \frac{s_i(t+) - A_{jk}(t)}{\sum_i Y_{ij}(t) w_i}$$

where $s_i(t+)$ is the observation's new state at time t. Any observation which is at risk contributes to only one row of the matrix derivative, the row corresponding to their starting state j, and $p_j(t-)$ times that row is added to that row of U. This "select my row" operation does not have a compact matrix formula, but is very simple computationally. Since each row of A must sum to 0, each row of the derivative sums to 0.

For the exponential form we have similarly

$$p(t) = p(t-)e^{A(t)}$$

$$U(t) = \frac{\partial p(t)}{w_i} = U(t-)e^{A(t)} + p(t-)\frac{\partial e^{A(t)}}{\partial w_i}$$

Derivatives of matrix exponential are a bit more involved, but follow directly from the definition:

$$\exp(A) = I + \sum_{i=1}^{\infty} A^i / i!$$

If we had a matrix B containing element-wise derivatives of A, then the derivative of the $A^3/3!$ term above would be $(BAA + ABA + AAB)/3!$. (The routines for the matrix exponential use more efficient algorithms for both the exp and its derivatives. They accept A and one or more B matrices, returning the exp

and derivative matrices.) Each derivative is a significant computation, and we would like to avoid separate computations of ∂w_i for all n observations in a study. Since A is a transition matrix each element of $\exp(A)$ is ≥ 0, each row sums to 1, and if there are no transitions from state j to k then the jk element is 0. With these and other considerations, no more than $s(s-1)$ different B matrices are ever needed.

A.4.3 Cox model

In all the formulas below we assume that the risks score for each subject have been centered at some covariate value z, i.e., $r_i = \exp((X_i - z)\beta)$.

Coefficients The derivative of the score statistic and $\hat{\beta}$ with respect to case weights are

$$\frac{\partial U}{\partial w_i} = \int_0^\infty (x_i - \overline{x}(s))dM_i(s) \tag{A.37}$$

$$\begin{aligned}\frac{\partial \hat{\beta}}{\partial w_i} &= \frac{\partial(\mathcal{I}^{-1}U)}{\partial w_i} \\ &= \mathcal{I}^{-1}\frac{\partial U}{\partial w_i} + \frac{\partial \mathcal{I}^{-1}}{\partial w_i}U \qquad \text{(A.38)} \\ &\approx \mathcal{I}^{-1}\frac{\partial U}{\partial w_i} \qquad \text{(A.39)}\end{aligned}$$

The right hand side of (A.37) is called the score residual. The inverse of the information matrix, $\mathcal{I}^{-1}$, is the usual asymptotic variance estimate of $\hat{\beta}$. The n by p matrix of derivatives of $\hat{\beta}$ is called the *dfbeta* matrix D, with ith row defined by (A.39). A robust variance for $\hat{\beta}$ is $D'W^2D$. The second term of (A.38) turns out to be small; it is a major nuisance to compute and so is ignored.

Cumulative hazard The predicted hazard and/or survival is always for a hypothetical subject; let z be the covariates for that subject. As above, we assume that the risk scores have been centered at z. The coefficients will be a matrix B with one column for each transition; for simplicity the derivation below use a vector of coefficients β and corresponding vector of transition specific risk scores r, (i.e., assumes you have selected the correct column from B). $Y_{ij}(t) = 1$ if observation i is currently in state j at time $t - 0$, i.e., is at risk for a transition from j to k, and $dN_{ijk}(t)$ is 1 if subject i has a $j : k$ transition

at time t.

$$\hat{\lambda}_{jk}(t;z) = \frac{\sum_i w_i dN_{ijk}(t)}{\sum_i w_i Y_{ij} t) r_i(t)} \tag{A.40}$$

$$\hat{\Lambda}_z(t) = \int_0^t \int_0^t \lambda_z(s)$$

$$\frac{\partial \hat{\lambda}_{jk}(t))}{\partial w_m} = \frac{dN_m(t)}{\sum_i w_i Y_{ij}(t) r_i(t)} - \frac{\sum_i w_i dN_i(t)}{\left(\sum_i w_i Y_{ij}(t) r_i(t)\right)^2} \frac{\partial \left(\sum_i w_i Y_{ij}(t) r_i(t)\right)}{\partial w_m} \tag{A.41}$$

$$= \frac{dM_{mjk}(t)}{\sum_i w_i Y_{ij}(t) r_i(t)} - \frac{\sum_i w_i dN_i(t)}{\left(\sum_i w_i Y_{ij}(t) r_i(t)\right)^2} \sum_i w_i Y_{ij}(t) \frac{\partial r_i(t)}{\partial w_m} \tag{A.42}$$

$$= A - B \tag{A.43}$$

$$\frac{\partial r_i(t)}{\partial w_m} = \frac{\partial \exp(\sum_{l=1}^p ((x_{il}(t) - z_l)\beta_j)}{\partial w_m}$$

$$= r_i(t) \sum_l (x_{il} - z_l) \frac{\partial \beta_l}{\partial w_m}$$

$$= r_i(t) \sum_l (X_{il}(s) - z_l')(D_{ml})' \tag{A.44}$$

$$B = \left(\frac{\sum_i w_i dN_i(t)}{\sum_i w_i Y_{ij}(t) r_i(t)} \right) \sum_{l=1}^p D_{ml} \frac{\sum_i Y_{ij}(t) w_i r_i(t)(x_{il} - z_l)}{\sum_i w_i Y_{ij}(t) r_i(t)} \tag{A.45}$$

$$= \hat{\lambda}_{jk}(t;z) \sum_{l=1}^p D_{ml}(\overline{x}_l(t) - z_l)$$

$$= \hat{\lambda}_{jk}(t;z) D_{m.}(\overline{x}(t) - z) \tag{A.46}$$

$$A - B = \frac{dN_m(t)}{\sum_i w_i Y_{ij}(t) r_i(t)} - \hat{\lambda}_{jk}(t;z) \left[\frac{Y_m(t)}{\sum_i w_i Y_{ij}(t) r_i(t)} + D_{m.}(\overline{x}(t) - z) \right] \tag{A.47}$$

Equation (A.40) is the definition of a hazard increment that we have used throughout. The step up to equation (A.43) reprise those for the Nelson-Aalen, leading to two terms A and B, the first of which is identical to that for the NA, the second having to do with the impact of each observation on $\hat{\beta}$. Step (A.44) recognizes the derivative with respect to β as the dfbeta matrix D defined earlier. The next step interchanges the order of summation for observations (i) and covariates (l), giving $\overline{x}$ and a final simplification.

A time-dependent reference z carries through in the above notation: at each jump in $\hat{\Lambda}$ there is a new covariate value driving the risk. Computer code is harder of course. However, the cumulative hazard and survival for a time-dependent covariate raise serious issues wrt what such curves actually mean, so there little call for this.

A.5 The IJ estimate of variance for the Kaplan-Meier

An interesting fact is that for the Kaplan-Meier, the variance based on the infinitesimal jackknife (IJ) is identical to the usual variance based on the Greenwood estimate. To make the notation in this section more compact, let $n(t) = \sum Y_i(t)w_i$ be the weighted number of subjects at risk at time t, $e(t) = \sum w_i dN_i(t)$ be the number of events at time t, and $d_1, d_2, \ldots$ be unique event times. The Greenwood and IJ variance estimates are

$$\begin{aligned} V_G(t) &= S^2(t) \sum_{d_j \le t} \frac{e(d_j)}{n(d_j)[n(d_j) - e(d_j)]} \\ &= S^2(t) g(t) \\ V_{IJ}(t) &= \sum_{i=1}^{n} u_i^2(t) \\ &= \sum_{i=1}^{n} \left(\frac{\partial S(t)}{\partial w_i} \right)^2 \end{aligned}$$

This equivalence is a surprise since formulas look nothing at all alike, but it was noticed in data examples that the two were identical. The following proof by induction is due to to Anne Eaton [32].

First:

$$\frac{\partial [1 - e(t)/n(t)]}{\partial w_i} = \frac{Y_i(t)e(t)}{n^2(t)} - \frac{dN_i(t)}{n(t)}$$

$$\begin{aligned} \sum_i w_i \left(\frac{\partial [1 - e(t)/n(t)]}{\partial w_i} \right)^2 &= \sum_i w_i Y_i(t) e^2(t)/n^4(t) + \sum_i w_i dN_i(t)/n^2(t) \\ &\quad - 2 \sum_i w_i Y_i(t) dN_i(t) e(t)/n^3(t) \qquad \text{(A.48)} \\ &= e^2(t)/n^3(t) + e(t)/n^2(t) - 2e^2(t)/n^3(t) \\ &= \frac{e(t)}{n^2(t)} \frac{n(t) - e(t)}{n(t)} \qquad \text{(A.49)} \end{aligned}$$

At the first event time d_1, $S(d_1) = 1 - e(d_1)/n(d_1)$, and from (A.49)

$$\begin{aligned} \sum_i u_i^2(d_1) &= \frac{e(d_1)}{n^2(d_1)} \frac{n(t) - e(t)}{n(t)} \frac{n(d_1) - e(d_1)}{n(d_1) - e(d_1)} \\ &= \left(\frac{n(d_1) - e(d_1)}{n(d_1)} \right)^2 \frac{e(d_1)}{n(d_1)[n(d_1) - e(d_1)]} \\ &= S^2(d_1) g(d_1) \qquad \text{(A.50)} \end{aligned}$$

which shows that the theorem is true at the first event time d_1.

Second: Assume that the theorem holds for event times $t < d_j$. Then

$$S(d_j) = S(d_{j-1})\frac{n(d_j) - e(d_j)}{n(d_j)} \tag{A.51}$$

$$u_i(d_j) = u_i(d_{j-1})\frac{n(d_j) - e(d_j)}{n(d_j)} + S(d_{j-1})\frac{\partial[1 - e(d_j)/n(d_j)]}{\partial w_i} \tag{A.52}$$

$$\sum_i u_i^2(d_j) \equiv \sum_i (a_i + b_i)^2 \tag{A.53}$$

$$\begin{aligned}\sum_i a_i^2 &= \sum_i u_i^2(d_{j-1})\left(\frac{n(d_j) - e(d_j)}{n(d_j)}\right)^2 \\ &= S^2(d_{j-1})g(d_{j-1})\left(\frac{n(d_j) - e(d_j)}{n(d_j)}\right)^2 \\ &= S^2(d_j)g(d_{j-1})\end{aligned} \tag{A.54}$$

$$\sum_i b_i^2 = S^2(d_{j-1})\left(\frac{n(d_j) - e(d_j)}{n(d_j)}\right)^2 \frac{e(d_j)}{n(d_j) - e(d_j)} \tag{A.55}$$

$$= S^2(d_j)\left(g(d_j) - g(d_{j-1})\right) \tag{A.56}$$

$$\sum_i a_i b_i = 0 \tag{A.57}$$

Equation (A.51) is the definition of the Kaplan-Meier, equation (A.52) follows from the product rule for derivatives, and (A.53) is shorthand to make the remaining lines fit on the page. Step (A.54) follows from the induction hypothesis, and (A.55) is a repeat of the steps for (A.50). Equation (A.57) follows from 3 observations

- $\frac{\partial 1 - e(d_j)/n(d_j)}{\partial w_i} = 0$ for all subjects i who are not at risk at time d_j
- $\sum_i \frac{\partial 1 - e(t)/n(t)}{\partial w_i} = 0$
- All subjects i who are at risk at time d_j share the same value for $u_i(d_{j-1})$.

The last statement reveals that this proof will not extend to a dataset with delayed entry.

A.6 IPC weights

Inverse probability of censoring weights appear in several contexts. An important aspect of these, often overlooked, is the handling of tied times. Consider a simple study with death as the endpoint and a particular death time at day 100, say. The analysis dataset will contain a death at day 100, and may also contain one or more censored observations at $t = 100$, as well as changes in one or more time-dependent covariates at $t = 100$. To sort this out we assume the following order: the deaths, then any censoring, then any covariate changes. The first of these is a reflection of the observation process itself: "censored on day 100" is the coding used for a patient who was last observed alive on day

100, i.e., their death time, whatever it is, must be > 100. Thus in the code censors happen after deaths, something that is baked in to every Kaplan-Meier or Cox model routine. The second condition, that covariate changes happen after everything else, is a consequence of the underlying mathematics, namely that the models are valid only if the covariates form what is known as a *predictable process*. In gambling parlance, you must place your bet before the roulette wheel is spun.

A.6.1 Inverse probability weight

The IPC weight is $1/G(t)$ where G is the censoring distribution. A common way to estimate this is with a Kaplan-Meier, with the 0/1 status variable reversed, and this approach is *almost* correct. The issue comes about with tied time values. Say we have a dataset with the following 4 observations for (time, status) of (31, 0), (52,0), (52,1), and (85,1). For computation of the ordinary KM of the event (death say), observations 2,3 and 4 are each considered to have been at risk for the event on day 52: censoring happens after the event. This agrees with how clinical data is gathered: a subject last observed to be alive on day 52 is coded as being censored on day 52; and we know that their death time is > 52. For computation of a Kaplan-Meier estimate $\hat{G}$ of the censoring distribution, however, the death at time 52 is not at risk for being censored at time 52: you cannot be "lost to followup" if you have already died. The risk set at time 31 will have all four obs, the risk set at time 52 should have only 2. One way to cause this to occur is to subtract a small value ϵ from each of the event times (but only those) before computing the Kaplan-Meier. An ϵ value of 1/2 the smallest difference between unique event times is a good default.

Asking whether this fix makes any practical difference is a fair question, and the answer is that normally the effect is small to negligible, and the "naive" reversed KM will be just fine. We are less charitable to those who are writing general software to be used by others; then the computation should be done correctly, for professional pride if no other reason.

A.6.2 Redistribute to the right

In the redistribute to the right algorithm this ordering is integral: when weights are redistributed, a subject who has an event at the concurrent time is not included in the recipient list. An alternate algorithm for the RTR process, and the one that is much more commonly used, is to give, at time t, each observation a weight of 0 if censored or the inverse probability of censoring (IPC) weight of $1/G(t)$ if uncensored, where G is a Kaplan-Meier estimate of the censoring distribution.

A common description of G is to "use the KM, with the status variable reversed", but this approach is correct only if there are no time points that

share a censoring and an event time. The correct algorithm has two important differences.

1. When computing $G(t)$ ensure that events are addressed before censoring times. One way to do this is to replace all event times t with $t - \epsilon$; then a computation with reversed status will retain the correct behavior.
2. When assigning weights at a given time t, those censored *before* t get a weight of zero, and all others a weight of $1/G(t + \epsilon)$. That is, use a left-continuous form of G rather than the default right-continuous form.

A reasonable value for ϵ is half the minimum spacing between any two unique time values in the dataset.

One feature of the RTR algorithm is that the sum of weights is constant over time, they are simply re-distributed. Use of the incorrect algorithm results in a sum the grows each time there is a shared event/censoring time.

Computing the Brier score

Create a "reversed status" survival curve $G(t)$ using a response of $(t-\epsilon, \delta = 0)$ for the events and $(t, 1)$ for the censored subjects. The small increment ϵ causes the correct accounting at time points with both an event and a censoring. At time 0 all subjects start with a weight of $1/n$, then going forward in time, at each event time s, in order,

1. Compute the weighted Brier score at time s
   $B(s) = \sum_{i=1}^{n} w_i(s)(y_i(s) - \hat{y}_i(s))^2$
2. Reset the weights of each subject who is still at risk ($t_i > s$) to $1/nG(s)$, and the weight of any subject censored at s to 0.

Doing the two steps in the proper order leads to left-continuity for G, which is important for the underlying theory. Using this approach the sum of the weights at any given time point will always be 1.

Many programs for the Brier score overlook the issue with ties, leading to a sum of weights that grows with time, just as for the RTR and IPC. The effect of this depends on the number of tied event/censoring values in the dataset, which is, thankfully often quite small. List some values?

A.7 Concordance

> "mille vie ducunt hominem per secula Romam" (a thousand roads lead men forever to Rome) Alain de Lille (1175)

For continuous data the concordance C between two variables x and y is defined as $P(y_i > y_j | x_i > x_j)$, i.e., the probability that the two variables share the same ordering for a randomly chosen pair of subjects i and j. For time-to-event data the concordance statistic has also become popular, modified to deal with censored data [49]. What is particularly interesting, however, is that there are close connections between concordance, two sample tests such as the Gehan-Wilcoxon statistic, and the Cox model score statistic. As such, there also turn out to be connections between suggestions for improvements, e.g., Uno's 2011 [122] modification of C is an echo of Schemper's 1992 [102] proposal for modifications of the Cox score statistic, which is itself a continuation of a 1972 Peto [89] modifcation of the Gehan-Wilcoxon test. This set of interconnections is not widely appreciated. (Several of them became clear in programming the survival package, i.e., deja vu moments we we realized that "I have seen this computation before".)

Though of interest, these historical connections are not central to message and purpose of the book, and recounting them in the body of the text would be a distraction. They do, however, inform our suggestions with respect to how the concordance statistic is best used and computed, and so are presented here in the appendix. We have largely omitted proofs as they can be found elsewhere.

A.7.1 Measures

We will focus on the concordance of a measured outcome y and the prediction $\hat{y}$ of that outcome. For continuous data, four common measurements of the concordance are Kendal's τ_a and τ_b, the Goodman-Kruska γ statistic, and Somers' d [1]; they differ only in how ties are handled. The four measures each lie between -1 and 1, similar to R^2.

Let A, D, T_x, T_y and T_{xy} be a count of the pairs that are concordant (or agree), discordant (disagree), tied on $\hat{y}$ (but not y), tied on y (but not $\hat{y}$), and

[1] or D_{xy}, depending on the reference

tied on both. $A + D + T_x + T_y + T_{xy} = n(n-1)/2$, the total number of pairs.

$$\tau_a = \frac{A - D}{n(n-1)/2} \tag{A.58}$$

$$\tau_b = \frac{A - D}{\sqrt{(A + D + T_x)(A + D + T_y)}} \tag{A.59}$$

$$\gamma = \frac{A - D}{A + D} \tag{A.60}$$

$$d = \frac{A - D}{A + D + T_x} \tag{A.61}$$

- Kendall's tau-a (A.58) is the most conservative; ties shrink the value of the statistic towards 0.
- The Goodman-Kruskal γ statistic (A.60) ignores ties in either y or $\hat{y}$.
- Somers' d (A.61) treats ties in y as incomparable and they are removed from the denominator.
- Kendall's tau-b (A.59) is similar to Somers' d, but treats y and $\hat{y}$ symmetrically.

Let $C = P(\hat{y}_i > \hat{y}_j | y_i > y_j)$, the fraction of time that a prediction turns out to be correct; $0 \leq C \leq 1$. Consider the following simple experiment

- Present pairs of subjects to an oracle — a statistical prediction rule, a domain expert such as an M.D., a local seer, whatever — and count the number of times that the oracle correctly predicts the order of the outcome.
- Pairs for which the outcome $y_i = y_j$ are not presented, as they would be uninformative with respect to the oracle's accuracy.
- Pairs for which the oracle cannot decide count as 1/2, e.g. tied values of the prediction rule.
- Define C = fraction correct.

For this approach of handling ties, it turns out that $C = (d + 1)/2$, a rescaled version of Somer's d.

This definition has two other equivalencies:

- If y is a 0/1 variable, then C = the area under the receiver operating curve (AUROC). (This is one of those identities that looks like it would have a 2 line proof, but takes a bit more work than that. Nevertheless it is well known.)
- When y is a survival time, then pairs for which the ordering of y_i and y_j cannot be ascertained are not used; this leads to Harrell's C_H.

For survival data, pairs of observations y that cannot be unabiguously ordered are treated as tied values. This would include the pair $(10+, 20)$, the first censored at time 10 and the second with an event at time 20: we do not know if observation 1 will have it's eventual event before or after time 20. The pair $(20+, 20)$ *is* ordered, however, since we know that the first observation's event happens sometime after $t = 20$.

A.7.2 Rank tests and the Cox model

For computation, start by sorting the data set by the observed survival response y. Then the numerator of Somers' d is

$$d = \sum_{i=1}^{n} \sum_{y_j > y_i} \text{sign}(\hat{y}_j - \hat{y}_i) \quad \text{(A.62)}$$

$$= \sum_{i=1}^{n} N_i \sum_{j \in R_i} \text{sign}(\hat{y}_j - \hat{y}_i) \quad \text{(A.63)}$$

$$= \sum_{i=1}^{n} \int \left[\sum_j Y_j(t) \text{sign}(\hat{y}_j(t) - \hat{y}_i(t)) \right] dN_i(t) \quad \text{(A.64)}$$

where N_i is the 0/1 indicator of censored/event and R_i is the risk set for the event (if any) for subject i. The transition from (A.62) to (A.63) follows first from our definition: if observation y_i is censored then any observations with a larger time are not comparable; the first sum need only include the events. The second insight is that within a set of tied times, the sum of the sign terms will be zero. Equation (A.64) writes this more carefully using counting process notation.

Antolini et al [7] made use of this identity to extend the C statistic to a model with crossing hazards. In this case one cannot make the simplifying assumption that $\eta_i > \eta_j$ ensures that thepredicted survival $S_j(t)$ for subject j will be greater than that for subject i, for any time t. Equation (A.64) only needs the ordering of predictions for i and y at the time of i's event. They apply this to a fit with a time-dependent coefficient.

Therneau and Watson [118] carried this one step further to rewrite (A.64) as the score statistic from a Cox model. This provides a direct connection between the concordance and many standard tests. Let D be the indices of the events (deaths) in the data set. Consider three intersecting facts

A. Let E be the set of data indices for the events (or deaths). For a single covariate x the score statistic for a time-weighted Cox model is

$$\sum_{i \in E} v(t_i)[x_i - \overline{x}(t_i)] \quad \text{(A.65)}$$

where t_i is the event time for observation i, $v(t)$ is a fixed time dependent weight function, and $\overline{x}(t)$ is the mean over all those at risk at time t. Normally $v(t) = 1$ for a Cox model. The variable x may also be time dependent.

B. When x is a 0/1 variable, then equation (A.65) is the Gehan-Wilcoxon test statistic when $v(t) = n(t) =$ the number at risk at time t. Other choices include the log-rank with $v(t) = 1$, the Peto-Wilcoxon with $v(t) = S(t)$, and several more.

C. Let $z(t)$ be a time dependent covariate defined as

$$z_i(t) = \frac{\sum_j Y_i(t) Y_j(t) sign(\hat{y}_i - \hat{y}_j)}{\sum_j Y_j(t)} \quad \text{(A.66)}$$

where *rmsign* is the sign function with value -1, 0, or 1. In words, for each observation i in the risk set at time t, z is the fraction of others whose prediction

is smaller, minus the fraction whose predicted value is larger; $r = (z+1)/2$ is a time dependent rank which lies in (0, 1). Then

$$C_H = \sum_{i \in D} n(t_i)[r_i - \overline{r}(t_i)] \tag{A.67}$$

is Harrell's C statistic. That is, C_H is equivalent to a Cox model score statistic, based on the time depedent *rank* of the prediction. (If there are tied values in $\hat{y}$ it is equivalent to the Cox model score statistic using the Breslow approximation.)

Three immediate consequences are that

1. Asymptotics for C_H now follow immediately from standard counting process arguments, in the same way as other Cox model results.
2. The concordance is well defined for time-dependent covariates.
3. The decades of consideration and debate over the "best" weighting $v(t)$ carry forward directly to the concordance.

The last of these is the most interesting. Peto and Peto [89] for instance noted that $n(t) \approx n(0)S(t)G(t)$, where S and G are the survival functions for death and censoring, respectively. They postulated that using $S(t)$ as the weight would result in a more stable test when the two treatment arms had different censoring patterns. Prentice [91] later showed this superiority for an empirical dataset, and indeed, implementations of the Gehan-Wilcoxon in the major statistics packages most often implement the Peto variant. We might even label equationeqconcord2 the "Gehan-Wilcoxon concordance" one with $v(t) = S(t)$ the Peto-Wilcoxon concordance, and etc.

Schemper [101] proposes a weight of $v(t) = S(t)/G(t)$ in the usual Cox model. The argument is that if proportional hazards does not hold, one would like an estimate that is consistent with respect to subject follow-up, that is, if one were to analyze the same study after t years of follow-up, and again after $t + s$ years, the coefficients from the two analyses will estimate the same quantity. This can be accomplished by using a time-dependent weight of $S(t-)/G(t-)$.

Uno et al. [122] essentially reprise the Schemper reasoning, and recommend a weight of $v(t) = n(t)/G^2(t-)$. Note that $n(t)/G^2(t-) \approx n(0)S(t-)/G(t-)$ by Peto's argument. Estimates using $\hat{G}$ and $\hat{S}$ turn out to be identical, actually (if G is calculated properly). They authors make the even stronger claim that the C statistic that will arise with complete follow-up of all the subjects is the "correct" target for estimation (we do not agree). Since in practice the hazard ratio almost invariably becomes closer to 1 over time in clinical studies, given sufficient follow-up, this means that the usual C_H statistic using $n(t)$ weighting will be larger than this asymptotic value.

A.7.3 Variance

The connection to the Cox model provides a natural measure of variance, namely the Cox model's second derivative or Hessian. This applies for any of the weighting choices, with or without time dependent covariates. The Cox model variance is correct under the null hypothesis, and so provides a valid test of concordance = 0.5. However, when C is far from 0.5 the resulting variance is too large, resulting in wider confidence intervals than needed.

An alternative is to compute a robust variance based on the infinitesimal jackknife. Rewriting equation (A.61) and adding case weights w, the numerator of Somers' d is

$$\begin{aligned} A &= \sum_{i \in E} v(t_i) w_i \sum_j w_j Y_j(t_i) I(t_j > t_i) \text{sign}(x_i - x_j) \\ U_k &= \frac{\partial A}{\partial w_k} \\ &= \sum_{i \in E} v(t_i) w_i Y_k(t_i) I(t_k > t_i) \text{sign}(x_i - x_k) + I(k \in E) v(t_k) \sum_j w_j Y_j(t_k) I(t_j > t_k) \text{sign}(x_k - x_j) \\ \text{var}(A) &= \sum_k U_k^2 \end{aligned}$$

A.7.4 Truncated values

We argue in section **??** that the approriate target of validation will often be a time-limited prediction. You might for instance compute the RMST over 2 years for "standard" treatment based on a fitted model for that treatment, using the result to decide on use of an experimental therapy. In that situation, the ability of the model to discriminate between a 3 year and a 5 year survival is immaterial, and it natural to compute the concordance on truncated data. That is $C(\min(y, 2), \hat{y})$, where all survival time > 2 are replaced by the censored value 2+ before doing the compuatation. Since the compuation only sums over pairs (i, j) where one survival time is known to be larger, this simple strategm effectively removes said 3 vs 5 year comparison from the calculation. In R, the concordance function has an optional `ymax` argument to make this particularly easy. Variance of the statistic carries through as before.

A more severe constraint is to dichomize the response as (survival $le 2$ years) versus (survival > 2 years). Since the response is now a 0/1 variable, without censoring, C is equal to the well known area under the receiver operating curve (AUROC). Our major problem with this comes from a career-long battle against the disease of "dichotomania"; the nearly relentless of our medical colleages to transform everything into black vs white. (We've sometimes joked about a new automobile made especially for medical researchers, with a single red light in place of the speedometer). The idea has been promoted under the label "time dependent AUROC", starting with a paper by Heagerty and Zheng [50] and with a large subsequent literature. This is a brilliant label,

making dichotomy appear to be a virtue rather than a vice. (It all feels like an own goal by the statistics team).

In its favor, the notion of "predicted 2 year survival" is easy to communicate to our audience, certainly easier than C or truncated C. As with all dichotomization there is a price to pay in terms of efficiency, the dichotomized C has larger variation. A primary technical issue is the "without censoring" qualifier above. What is to be done with the subject censored at 1.5 years, when creating the 0/1 variable for 2 year survival? One solution is to apply the RTTR algorithm to all those censored before time 2, redistributiong their case weight, then compute a weighted concordance. Figure **??** shows the truncated and dichomized C statistic at multiple time points for a model fit to the Rotterdam data, assessed on the GBSG data set as validation, along with 1 SE error bars. For the truncated data both the Harrell and Uno weightings are shown, the dichomized statistic has dealt with censoring in another way. [This figure is in the validation chapter.]

A.7.5 Synthetic measures

Göen and Heller [44] take a different approach to the censoring issue. They note that if the proportional hazards model holds, then given two subjects with risk scores η_i and η_j, the probability that $t_i > t_j$ is $\exp(\eta_j - \eta_i)$. Looking ahead from time 0, one can compute the expected concordance of the model $E(C)$ directly using only the covariate matrix X and the estimated coefficient $\hat{\beta}$. This completely sidesteps censoring.

Sort the data in order of the estimated risk score $\hat{\eta}$, then the estimate is

$$C_G = \sum_{i>j} \frac{1}{1 + e * \eta_j - \eta_i} / [n(n-1)/2] \tag{A.68}$$

The significant downside to this approach is that we get an estimate of how well the PH model *would* predict, in some future dataset that has complete follow-up and exactly follows the model, i.e., proportional hazards over all time, perfect linearity wrt covariates, and no outliers. This is, however, not an evaluation of how well the model *actually* works for the data set at hand.

A.8 Dates and roundoff error

The raw data for studies often has dates of entry and exit from a study, which means that the natural unit of time for follow-up is days. Or we should say that is the form which is natural to the computer, since software packages like R, SAS, and STATA can all subtract dates directly. Users will often be more comfortable working in years, however, either age or years since entry for instance. It is not widely appreciated that this can lead to some subtle round off errors.

As an example, consider two subjects, both born on 1973-3-10, the first

enrolled on 1998-09-13 and followed until 12-04, the second enrolled on 1998-09-17 and followed until 12-08; both have 81 days of followup. However, if we compute age at enrollment as (entry date - birth date)/365.25, and similarly for age at last follow-up, the simple difference (follow-up age - entry age) will *not* be exactly the same for the two subjects. This is due to the fact that 365.25 does not have an exact representation in binary,

For parametric models this tiny error has no impact whatsoever, but for the Kaplan-Meier and Cox model exact ties are treated differently, since they determine which observations are or are not in a given risk set. Users can be surprised when a change from days to years leads to a different result. (Some software is aware of this and tries to avoid the problem, most is not.)

A.9 Approximating a Cox model

The Cox model can be approximated using Poisson regression, a trick that was well known when Cox models were not yet common in the major software packages [128, 67]. The coefficients and standard errors for the Poisson regression, which uses the number of events for each person as the y variable, are usually quite close to those of the Cox model, which is focused on a censored time value as the response. In fact, if the baseline hazard of the Cox model $\lambda_0(t)$ is assumed to be constant over time, the Cox model is equivalent to Poisson regression.

One non-obvious feature of a Poisson fit is the use of an `offset` term. This is based on a clever "sleight of hand", which has its roots in the fact that a Poisson likelihood is based on the number of events (y), but that we normally want to model not the number but rather the *rate* of events (λ). Then

$$\begin{aligned} E(y_i) &= \lambda_i t_i \\ &= \left(e^{X_i\beta}\right) t_i \\ &= e^{X_i\beta + \log(t_i)} \end{aligned} \tag{A.69}$$

We see that treating the log of time as another covariate, with a known coefficient of 1, correctly transforms from the hazard scale to the total number of events scale. An `offset` in glm models is exactly this, a covariate with a fixed coefficient of 1.

The hazard rate in a Poisson model is traditionally modeled as $\exp(X\beta)$ (i.e. the inverse link $f(\eta) = e^{\eta}$) rather than the linear form $X\beta$, for essentially the same reason that it is modeled that way in the Cox model: it guarantees that the hazard rate (the expected value given the linear predictors) is positive.

Example

In order to better understand the underlying hazard rates, start by plotting the cumulative hazard for the dataset, and approximate it with a set of connected line segments. The kidney cancer data, for instance, is moderately well approximated using cutpoints at 45 and 500 days, as shown in Figure A.1.

The Poisson regression approximation can then be obtained by breaking the time scale into intervals based on these cutpoints and fitting a Poisson model with one intercept per interval. We see that the coefficients and standard errors are very similar (Table A.2, row 2).

If each unique death time is made into its own interval the Cox result can be duplicated exactly. One more correction is needed for perfect agreement, which is to toss away any partial contributions. If there were unique death times at 100 and 110 days, for instance, and a subject were censored at time 103, those 3 days are not counted in the in the 100–110 interval. If there is someone with follow-up after the last event, that is removed as well. After removal, everyone in the same interval will have the same number of days at risk, which means that an offset correction is no longer needed.

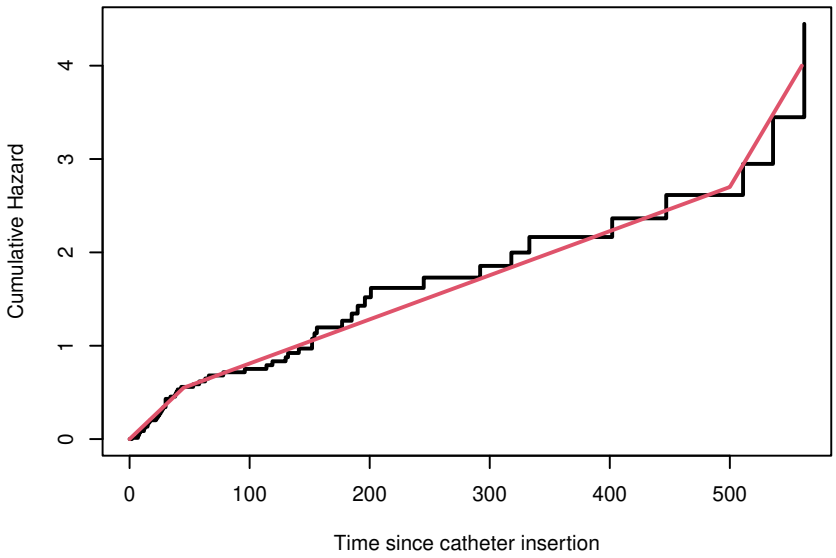

Figure A.1 *Cumulative hazard rates and approximations of the rate in three time periods using the kidney dataset with cutpoints at 45 and 500 days.*

	age	sex
Cox	0.0022 (0.0092)	-0.8210 (0.2987)
poisson1	0.0032 (0.0093)	-0.7512 (0.2943)
poisson2	0.0022 (0.0092)	-0.8210 (0.2987)
binomial	0.0028 (0.0095)	-0.8993 (0.3138)

Table A.2 *Comparison of Cox, poisson and binomial models of the kidney data predicting survival using the covariates age and sex. The poisson1 results are based on breaking follow-up into three intervals (cutpoints at 45 and 500 days) and the poisson2 results use intervals based on each unique death time. The binomial model uses the same data as was used for poisson2.*

The Poisson coefficients now exactly match those of the Cox model (Table A.2, row 3). Almost all of the intervals have only a single event, i.e., both the event count and the event rate are low, which is the case in which the binomial and Poisson distributions closely approximate each other. Consequently, the last model in Table A.2 shows that binomial fits are also effective. This computational trick can be particularly useful in contexts where there is readily available code for binomial outcomes but time-to-event models are lacking, e.g., machine learning.

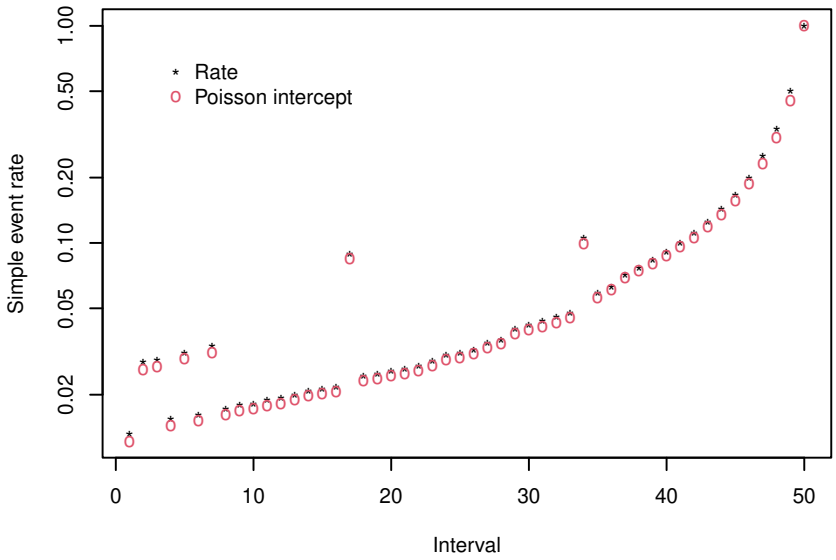

Figure A.2 *Poisson intercepts and event rates separated by intervals where events occurred using the kidney dataset.*

Pre-centering data

We can make the connection between the Cox, poisson, and binomial models easier to exploit by pre-centering the data. The plot shows that when this is done, the intercepts are very close to the simple event rate for each interval of (number of events) / (number of observations). We can add this variable to the model as an offset.

	age	sex
Cox	0.00218 (0.00922)	-0.82100 (0.29872)
Poisson, multiple intercepts	0.00218 (0.00922)	-0.82100 (0.29872)
Poisson, offset interval rate	0.00217 (0.00920)	-0.81845 (0.29757)
Poisson, no offset or intercept	0.00381 (0.00903)	-0.78852 (0.29063)
Binomial, offset interval rate	0.00278 (0.00950)	-0.89619 (0.31210)
Binomial, no offset	0.00403 (0.00928)	-0.83705 (0.30074)

Table A.3 *Comparison of Cox, poisson and binomial models of the kidney data predicting survival using the covariates age and sex. The data has been split into intervals based on event times and the data has been pre-centered. Coefficients and standard errors are shown for the different models.*

The coefficients from the different models, as shown in Table A.3, illustrate that adding an approximate per-interval intercept via the offset term gives a very close approximation to the `coxph` fit with the Poisson and a reasonable