Validation

Terry M Therneau

Dec 2015

1 Introduction

> `When I use a word,' Humpty Dumpty said, in rather a scornful tone, `it means just what I choose it to mean - neither more nor less.'
>
> `The question is,' said Alice, `whether you can make words mean so many di erent things.'
>
> `The question is,' said Humpty Dumpty, `which is to be master - that's all.'
>
> -Lewis Caroll, *Through the Looking Glass*

\Validatation" has become a Humpty-Dumpty word: it is used for many di erent things in scienti c research that it has become essentially meaningless without further clari cation. One of the more common meanings assigned to it in the software realm is \repeatability", i.e. that a new release of a given package or routine will give the same results as it did the week before. Users of the software often assume the word implies a more rigorous criterion, namely that the routine gives *correct* answers.

Validation of the latter type is rare, however; working out formally correct answers is boring, tedious work. This note contains a set of examples of this latter type. Although the data sets are very simple, the examples have proven extremely useful in debugging the methods, not least because all the intermediate steps of each calculation are transparent, and have been incorporated into the formal test suite for the survival package as the les `book1.R`, `book2.R`, etc. in the `tests` subdirectory. They also continue to be a resource for package's defence: I have been told multiple times that some person or group cannot use R in their work because \SAS is validated" while R is not. The survival package passes all of the tests below and SAS passes some but not all of them.

It is my hope that the formal test cases will be a resource for developers on multiple platforms. Portions of this work were included as an appendix in the textbook of Therneau and Grambsch [3] precisely for this reason.

2 Basic formulas

All these examples have a single covariate. Let x_i be the covariate for each subject, $r_i = exp(x_i)$ the risk score for the subject, and w_i the case weight, if any. Let $Y_i(t)$ be 1 if subject i is at risk at time t and 0 otherwise, and $dN_i(t)$ be the death indicator which is 1 if subject i has an event

at time t. At each death time we have the following quantities:

$$LPL(t) = \frac{\sum_i dN_i(t)\log(r_i)}{\log\left(\sum_i Y_i(t)w_i r_i\right)} \tag{1}$$

$$\bar{x}(t) = \frac{\sum_i Y_i(t)w_i r_i x_i}{\sum_i Y_i(t)w_i r_i} \tag{2}$$

$$U(t) = \sum_i dN_i(t)(x_i - \bar{x}(t)) \tag{3}$$

$$H(t) = \frac{\sum_i Y_i(t)(x_i - \bar{x}(t))^2}{\sum_i Y_i(t)w_i r_i} \tag{4}$$

$$= \frac{\sum_i Y_i(t)x_i^2}{\sum_i Y_i(t)w_i r_i} - \bar{x}(t)^2 \tag{5}$$

$$\Delta(t) = \frac{\sum_i w_i dN_i(t)}{\sum_i Y_i(t)w_i r_i} \tag{6}$$

The denominator for all the sums is the weighted number of subjects who are at risk, LPL is the contribution to the log partial likelihood at time t, and $\bar{x}$ and H are the weighted mean and variance of the covariate x at each time. The sum of $H(t)$ over the death times is the second derivative of the LPL, also known as the Hessian matrix. U is the contribution to the first derivative of the LPL at time t and Δ is the increment in the baseline hazard function.

3 Test data 1

This data set of $n = 6$ subjects has a single 0/1 covariate x. There is one tied death time, one time with both a death and a censored observation, one with only a death, and one with only censoring. (This is as small as a data set can be and still cover these four important cases.) Let $r = \exp(\beta)$ be the risk score for a subject with $x = 1$; the risk score is $\exp(0) = 1$ for those with $x = 0$. Table 1 shows the data set along with the mean and increment to the hazard at each time point.

			$\bar{x}(t)$		$\hat{d}_0(t)$	
Time	Status	x	Breslow	Efron	Breslow	Efron
1	1	1	$r/(r+1)$	$r/(r+1)$	$1/(3r+3)$	$1/(3r+3)$
1	0	1				
6	1	1	$r/(r+3)$	$r/(r+3)$	$1/(r+3)$	$1/(r+3)$
6	1	0	$r/(r+3)$	$r/(r+5)$	$1/(r+3)$	$1/(r+5)$
8	0	0				
9	1	0	0	0	1	1

Table 1: Test data 1

3.1 Breslow estimates

The log partial likelihood (LPL) has a term for each event; each term is the log of the ratio of the score for the subject who had an event over the sum of scores for those who did not. The LPL, first derivative U of the LPL and second derivative (or Hession) H are:

$$\begin{aligned} LPL &= f \quad \log(3r+3)g + f \quad \log(r+3)g + f0 \quad \log(r+3)g + f0 \quad 0g \\ &= 2 \quad \log(3r+3) \quad 2\log(r+3): \end{aligned}$$

$$\begin{aligned} U &= 1 \quad \frac{r}{r+1} + 1 \quad \frac{r}{r+3} + 0 \quad \frac{r}{r+3} + (0 \quad 0) \\ &= \frac{r^2+3r+6}{(r+1)(r+3)}: \end{aligned}$$

$$\begin{aligned} H &= \left(\frac{r}{r+1} \quad \frac{r}{r+1}^2\right) + 2\left(\frac{r}{r+3} \quad \frac{r}{r+3}^2\right) + (0 \quad 0) \\ &= \frac{r}{(r+1)^2} + \frac{6r}{(r+3)^2}: \end{aligned}$$

(For a 0/1 covariate the variance formula (5) simplifies to $\bar{x} \quad \bar{x}^2$, but only in that case. We used this fact above.)

The following function computes these quantities.

```
> breslow1 <- function(beta) {
      # first test data set, Breslow approximation
      r = exp(beta)
      lpl = 2*beta - (log(3*r +3) + 2*log(r+3))
      U   = (6+ 3*r - r^2)/((r+1)*(r+3))
      H   =  r/(r+1)^2 + 6*r/(r+3)^2
      c(beta=beta, loglik=lpl, U=U, H=H)
  }
> beta <- log((3 + sqrt(33))/2)
> temp <- rbind(breslow1(0), breslow1(beta))
> dimnames(temp)[[1]] <- c("beta=0", "beta=solution")
> temp
                   beta    loglik U         H
beta=0         0.000000 -4.564348 1 0.6250000
beta=solution 1.475285 -3.824750 0 0.6341681
```

The maximum partial likelihood occurs when $U(\ ) = 0$, namely $r^2 \quad 3r \quad 6 = 0$. Using the usual formula for a quadratic equation gives $r = (1=2)(3 + \sqrt{33})$ and $\hat{} = \log(r) \quad 1{:}475285$. The above call to `breslow1` verifies that the first derivative is zero at this point.

Newton{Raphson iteration has increments of $H^{-1}U$. Starting with the usual initial estimate of $= 0$, the first iteration is $= 8=5$ and further ones are shown below.

```
> iter <- matrix(0, nrow=6, ncol=4,
                 dimnames=list(paste("iter", 0:5),
                               c("beta", "loglik", "U", "H")))
> # Exact Newton-Raphson
> beta <- 0
> for (i in 1:6) {
      iter[i,] <- breslow1(beta)
      beta <- beta + iter[i,"U"]/iter[i,"H"]
  }
> print(iter, digits=10)
              beta       loglik                U
iter 0 0.000000000 -4.564348191  1.000000000e+00
iter 1 1.600000000 -3.829619615 -7.758917124e-02
iter 2 1.472723532 -3.824751586  1.624953335e-03
iter 3 1.475283961 -3.824749505  6.050799925e-07
iter 4 1.475284915 -3.824749505  8.413997922e-14
iter 5 1.475284915 -3.824749505 -8.970147038e-17
                  H
iter 0 0.6250000000
iter 1 0.6096112864
iter 2 0.6346411804
iter 3 0.6341683192
iter 4 0.6341681428
iter 5 0.6341681428
> # coxph fits
> test1 <- data.frame(time=  c(1, 1, 6, 6, 8, 9),
                      status=c(1, 0, 1, 1, 0, 1),
                      x=     c(1, 1, 1, 0, 0, 0))
> temp <- matrix(0, nrow=6, ncol=4,
                   dimnames=list(1:6, c("iter", "beta", "loglik", "H")))
> for (i in 0:5) {
      tfit <- coxph(Surv(time, status) ~ x, data=test1,
                    ties="breslow", iter.max=i)
      temp[i+1,] <- c(tfit$iter, coef(tfit), tfit$loglik[2], 1/vcov(tfit))
  }
> temp
  iter     beta    loglik         H
1    0 0.000000 -4.564348 0.6250000
2    2 1.600000 -3.829620 0.6096113
3    3 1.472724 -3.824752 0.6346412
4    4 1.475284 -3.824750 0.6341683
5    4 1.475285 -3.824750 0.6341681
6    4 1.475285 -3.824750 0.6341681
```

The `coxph` routine declares convergence after 4 iterations for this data set, so the last two calls with `iter.max` of 4 and 5 give identical results.

The martingale residuals are de ned as $O - E$ = observed - expected, where the observed is the number of events for the subject (0 or 1) and E is the expected number assuming that the model is completely correct. For the rst death all 6 subjects are at risk, and the martingale formulation views the outcome as a lottery in which the subjects hold r, r, r, 1, 1 and 1 tickets, respectively. The contribution to E for subject 1 at time 1 is thus $r/(r + 3)$. Carrying this forward the residuals can be written as simple function of the cumulative baseline hazard $_0(t)$, the Nelson cumulative hazard estimator with case weights of $w_i r_i$; this is shown in the `Breslow' column of table 1. (Also known as the Aalen estimate, Breslow estimate, and all possible combinations of the three names.) Then the residual can be written as

$$M_i = \quad _i \quad \exp(x_i \quad)\hat{}(t_i) \tag{7}$$

Each of the two subjects who die at time 6 are credited with the full hazard increment at time 6. Residuals at $= 0$ and $\hat{}$ are shown in the table below.

Subject	$_0$	$M(0)$	$M(\hat{})$
1	$1/(3r + 3)$	5/6	0.728714
2	$1/(3r + 3)$	1/6	-0.271286
3	$1/(3r + 3) + 2/(r + 3)$	1/3	-0.457427
4	$1/(3r + 3) + 2/(r + 3)$	1/3	0.666667
5	$1/(3r + 3) + 2/(r + 3)$	2/3	-0.333333
6	$1/(3r + 3) + 2/(r + 3) + 1$	2/3	-0.333333

The score statistic U can be written as a two way sum involving the covariate(s) and the martingale residuals

$$U = \sum_{i=1}^{} \int [x_i \quad \bar{x}(t)]dM_i(t) \tag{8}$$

The martingale residual M

residuals, one per event, rather than one per death time as this has proven to be more useful for plots and other downstream computations.

In the multivariate case there will be a matrix like the above for each covariate. Let L be the n by p matrix made up of the collection of row sums where n is the number of subjects and p is the number of covariates, this is the matrix of score residuals. The dfbeta residuals are the n by p matrix $D = LH^{-1}$; H has been de ned above for this data set. D is an approximate measure of the in uence of each observation on the solution vector. Similarly, the scaled Schoenfeld residuals are the (number of events) by p matrix obtained by multiplying the Schoenfeld residuals by H^{-1}.

As stated above there is a close connection between the Nelson{Aalen estimate estimate of cumulative hazard and the Breslow approximation for ties. The baseline hazard is shown as the column $_0$ in table 1. The estimated hazard for a subject with covariate x_i is $_i(t) = \exp(x_i\ )\ _0(t)$ and the survival estimate for the subject is $S_i(t) = \exp(-\ _i(t))$. The variance of the cumulative hazard is the sum of two terms. Term 1 is a natural extension of the Nelson{Aalen estimator to the case where there are weights. It is a running sum, with an increment at each death of $1=(\sum Y_i(t)r_i(t))^2$. For a subject with covariate x_i this term is multiplied by $[\exp(x_i\ )]^2$. The second term is $cH^{-1}c'$, where H is the information matrix of the Cox model and c is a vector. The second term accounts for the fact that the weights themselves have a variance; c is the derivative of $S(t)$ with respect to and can be formally written as

$$\exp(x\ )\ \int_0^t (x(s)\ \ x_i)d\hat{}_0(s):$$

This can be recognized as 1 times the score residual process for a subject with x_i as covariates and no events; it measures leverage of a particular observation on the estimate of . It is intuitive that a small score residual | an observation whose covariates has little in uence on | results in a small added variance; that is, has little in uence on the estimated survival.

Time	Term 1
1	$1=(3r + 3)^2$
6	$1=(3r + 3)^2 + 2=(r + 3)^2$
9	$1=(3r + 3)^2 + 2=(r + 3)^2 + 1=1^2$

Time	c
1	$(r=(r + 1))\quad 1=(3r + 3)$
6	$(r=(r + 1))\quad 1=(3r + 3) + (r=(r + 3))\quad 2=(r + 3)$
9	$(r=(r + 1))\quad 1=(3r + 3) + (r=(r + 3))\quad 2=(r + 3) + 0\quad 1$

For = 0, $x = 0$:

Time	Variance			
1	1/36	+ 1:6	$(1=12)^2$	= 7/180
6	(1/36 + 2/16)	+ 1:6	$(1=12 + 2=16)^2$	= 2/9
9	(1/36 + 2/16 + 1)	+ 1:6	$(1=12 + 2=16 + 0)^2$	= 11/9

For = 1:4752849, $x = 0$

Time	Variance		
1	0.0038498	+ .004021	= 0.007871
2	0.040648	+ .0704631	= 0.111111
4	1.040648	+ .0704631	= 1.111111

3.2 Efron approximation

The Efron approximation [1] di ers from the Breslow only at day 6, where two deaths occur. A useful way to view the approximation is to recast the problem as a lottery model. On day 1 there were 6 subjects in the lottery and 1 ticket was drawn, at which time the winner became ineligible for further drawings and withdrew. On day 6 there were 4 subjects in the drawing (at risk) and two tickets (deaths) were drawn. The Breslow approximation considers all four subjects to be eligible for both drawings, which implies that one of them could in theory have won both, that is, died twice. This is of clearly impossible. The Efron approximation treats the two drawings on day 6 as sequential. All four living subjects are at risk for the rst of them, then the winner is withdrawn. Three subjects are eligible for the second drawing, either subjects 3, 5, and 6 or subjects 2, 5, and 6, but we do not know which. In some sense then, subjects 3 and 4 each have \.5 probability" of being at risk for the second event at time 6. In the computation, we treat the two deaths at time 6 as two separate times (two terms in the loglik), with subjects 3 and 4 each having a case weight of 1/2 for the second one. The mean covariate for the second event is then

$$\frac{1 \quad r=2 + 0 \quad 1=2 + 0 \quad 1 + 0 \quad 1}{r=2 + 1=2 + 1 + 1} = \frac{r}{r + 5}$$

and the main quantities are

$$\begin{aligned} LL &= f \quad \log(3r + 3)g + f \quad \log(r + 3)g + f0 \quad \log(r=2 + 5=2)g + f0 \quad 0g \\ &= 2 \quad \log(3r + 3) \quad \log(r + 3) \quad \log(r=2 + 5=2) \end{aligned}$$

$$\begin{aligned} U &= 1 \quad \frac{r}{r + 1} + 1 \quad \frac{r}{r + 3} + 0 \quad \frac{r}{r + 5} + (0 \quad 0) \\ &= \frac{r^3 + 23r + 30}{(r + 1)(r + 3)(r + 5)} \end{aligned}$$

$$\begin{aligned} I &= \left(\frac{r}{r + 1} \quad \frac{r}{r + 1}\right)^2 + \left(\frac{r}{r + 3} \quad \frac{r}{r + 3}\right)^2 \\ &\quad + \left(\frac{r}{r + 5} \quad \frac{r}{r + 5}\right)^2 : \end{aligned}$$

The solution corresponds to the one positive root of

event at time 6 subjects 3 and 4 have a weight of 1/2, the total number of tickets is $(r + 5)/2$ and the consequent increment in the cumulative hazard is $2/(r + 5)$. This $= 0$ this calculation is equivalent to the Fleming-Harrington [2] estimate of cumulative hazard. Subjects 3 and 4 receive 1/2 of this second increment to E and subjects 5 and 6 the full increment. Efron [1] did not discuss residuals so did not investigate this aspect of the approximation, we nevertheless sometime refer to this using combinations of Fleming, Harrington, Efron in the same way as the Nelson-Aalen-Breslow estimate. The martingale residuals are

Subject	M_i				
1	1	$r/(3r + 3)$			
2	0	$r/(3r + 3)$			
3	1	$r/(3r + 3)$	$r/(r + 3)$	$r/(r + 5)$	
4	1	$1/(3r + 3)$	$1/(r + 3)$	$1/(r + 5)$	
5	0	$1/(3r + 3)$	$1/(r + 3)$	$2/(r + 5)$	
6	0	$1/(3r + 3)$	$1/(r + 3)$	$2/(r + 5)$	1

giving residuals at $= 0$ of 5/6, -1/6, 5/12, 5/12, -3/4 and -3/4.

The matrix de ning the score and Schoenfeld residuals has the same rst column (time 1) and last column as before, with the following contributions at time 6.

	Time								
Subject	6 (rst)				6 (second)				
1	0				0				
2	0				0				
3	1	$\frac{r}{r+3}$	1	$\frac{r}{r+3}$	1	$\frac{r}{r+5}$	1	$\frac{2r}{r+5}$	=2
4	0	$\frac{r}{r+3}$	0	$\frac{1}{r+3}$	0	$\frac{r}{r+5}$	1	$\frac{2}{r+5}$	=2
5	0	$\frac{r}{r+3}$	0	$\frac{1}{r+3}$	0	$\frac{r}{r+5}$	0	$\frac{2}{r+5}$	
6	0	$\frac{r}{r+3}$	0	$\frac{1}{r+3}$	0	$\frac{r}{r+5}$	0	$\frac{2}{r+5}$	

The score residuals at $= 0$ are 5/12, -1/12, 55/144, -5/144, 29/144 and 29/144.

It is an error to generate residuals for the Efron method by using formula (7), which was derived from the Breslow approximation. It is clear that some packages do exactly this, however, which can be veri ed using formulas from above. (Statistical forensics is another use for our results.) What are the consequences of this? On a formal level the resulting \martingale residuals" no longer have an expected value of 0 and thus are not martingales, so one loses theoretical backing for derived plots or statistics. The score, Schoenfeld, dfbeta and scaled Schoenfeld residuals are based on the martingale residual so su er the same loss. On a practical level, when the fraction of ties is small it is quite often the case that $\hat{}$ is nearly the same when using the Breslow and Efron approach. We have normally found the correct and ad hoc residuals to be similar as well in that case, su ciently so that explorations of functional form (martingale residuals), leverage and robust variance (dfbeta) and proportional hazards (scaled Schoenfeld) led to the same conclusions. This will not hold when there are a moderate to large number of ties.

The variance formula for the baseline hazard function in the Efron case is evaluated the same way as before, as the sum of (hazard increment)^2, treating a tied death as multiple separate

hazard increments. In term 1 of the variance, the variance increment at time 6 is now $1=(r+3)^2 + 4=(r+5)^2$ rather than $2=(r+3)^2$. The increment to d at time 6 is $(r=(r+3))\ \ 1=(r+3) + (r=(r+5))\ \ 2=(r+5)$. (Numerically, the result of this computation is intermediate between the Nelson{Aalen variance and the Greenwood variance used in the Kaplan{Meier.)

For $\ = 0$, $x = 0$, let $v = H^{-1} = 144=83$.

Time	Variance	
1	1/36	
	$+ v(1=12)^2$	$= 119/2988$
6	$(1/36 + 1/16 + 4/25)$	
	$+ v(1=12 + 1=16 + 1=18)^2$	$= 1996/6225$
9	$(1/36 + 1/16 + 4/25 + 1)$	
	$+ v(1=12 + 1=16 + 1=18 + 0)^2$	$= 8221/6225$

For $\ = 1{:}676857$, $x = 0$.

Time	Variance	
1	0.00275667 + .00319386	= 0.0059505
2	0.05445330 + .0796212	= 0.134075
4	1.05445330 + .0796212	= 1.134075

3.3 Exact partial likelihood

Returning to the lottery analogy, for the two deaths at time 6 the exact partial likelihood computes the direct probability that those two subjects would be selected given that a pair will be chosen. The numerator is $r_3 r_4$, the product of the risk scores of the subjects with an event, and the denominator is the sum over all 6 pairs who could have been chosen: $r_3 r_4 + r_3 r_5 + r_3 r_6 + r_4 r_5 + r_4 r_6 + r_5 r_6$. (If there were 10 tied deaths from a pool of 60 available the sum will have over 75 billion terms, each a product of 10 values; a truly formidable computation!) In our case, three of the four subjects at risk at time 6 have a risk score of $\exp(0x) = 1$ and one a risk score of r, and the denominator is $r + r + r + 1 + 1 + 1$.

$$\begin{aligned} LL &= f \quad \log(3r+3)g + f \quad \log(3r+3)g + f0 \quad 0g \\ &= 2f \quad \log(3r+3)g{:} \end{aligned}$$

$$\begin{aligned} U &= \quad 1 \quad \frac{r}{r+1} \quad + \quad 1 \quad \frac{r}{r+1} \quad + (0 \quad 0) \\ &= \quad \frac{2}{r+1}{:} \end{aligned}$$

$$H = \frac{2r}{(r+1)^2}{:}$$

The solution $U(\ ) = 0$ corresponds to $r = \ 1$, with a loglikelihood that asymptotes to $2\log(3) = 2.1972$. The Newton{Raphson iteration has increments of $(r+1)=r$ leading to the following iteration for $\hat{}$:

```
> temp <- matrix(0, 8, 3)
> dimnames(temp) <- list(paste0("iteration ", 0:7, ':'), c("beta", "loglik", "H"))
> bhat <- 0
> for (i in 1:8) {
      r <- exp(bhat)
      temp[i,] <- c(bhat,  2*(bhat - log(3*r +3)), 2*r/(r+1)^2)
      bhat <- bhat + (r+1)/r
  }
> round(temp,3)
              beta loglik     H
iteration 0: 0.000 -3.584 0.500
iteration 1: 2.000 -2.451 0.210
iteration 2: 3.135 -2.282 0.080
iteration 3: 4.179 -2.228 0.030
iteration 4: 5.194 -2.208 0.011
iteration 5: 6.200 -2.201 0.004
iteration 6: 7.202 -2.199 0.001
iteration 7: 8.202 -2.198 0.001
```

The Newton-Raphson iteration quickly settles down to addition of a constant increment to ^ at each step while the partial likelihood approaches an asymptote: this is a fairly common case when the Cox MLE is in nite. A solution at ^ = 10 or 15 is hardly di erent in likelihood from the true maximum, and most programs will stop iterating around this point. The information matrix, which measures the curvature of the likelihood function at , rapidly goes to zero as grows.

It is di cult to describe a satisfactory de nition of the expected number of events for each subject and thus a de nition of the proper martingale residual for the exact calculation. Among other things it should lead to a consistent score residual, i.e., ones that sum to the total score statistic U

$$L_i = \int (x_i - \bar{x}(t))dM_i(t)$$

$$\sum L_i = U$$

The residuals de ned above for the Breslow and Efron approximations have this property, for instance. The exact partial likelihood contribution to U for a set of set of k tied deaths, however, is a sum of all subsets of size k; how would one partition this term as a simple sum over subjects?

The exact partial likelihood is infrequently used and examination of post t residuals is even rarer. The survival package (and all others that I know of) takes the easy road in this case and uses equation (7) along with the Nelson-Aalen-Breslow hazard to form residuals. They are certainly not correct, but the viable options were to use this, the Efron residuals, or print an error message.

At ^ = *1* the Breslow residuals are still well de ned. Subjects 1 to 3, those with a covariate of 1, experience a hazard of r=(3r + 3) = 1=3 at time 1. Subject 3 accumulates a hazard of 1/3 at time 1 and a further hazard of 2 at time 6. The remaining subjects are at an in nitely lower risk during days 1 to 6 and accumulate no hazard then, with subject 6 being credited with 1

Time	Status	x	Number at Risk	$\bar{x}$	$\hat{d}$
(1,2]	1	1	2	$r/(r+1)$	$1/(r+1)$
(2,3]	1	0	3	$r/(r+2)$	$1/(r+2)$
(5,6]	1	0	5	$3r/(3r+2)$	$1/(3r+2)$
(2,7]	1	1	4	$3r/(3r+1)$	$1/(3r+1)$
(1,8]	1	0	4	$3r/(3r+1)$	$1/(3r+1)$
(7,9]	1	1	5	$3r/(3r+2)$	$2/(3r+2)$
(3,9]	1	1			
(4,9]	0	1			
(8,14]	0	0	2	0	0
(8,17]	0	0	1	0	0

Table 2: Test data 2

unit of hazard at the last event. The residuals are thus 1 1/3 = 2/3, 0 1/3, 1 7/3 = 4/3, 1 0, 0, and 0, respectively, for the six subjects.

4 Test data 2

This data set also has a single covariate, but in this case a (start, stop] style of input is employed. Table 2 shows the data sorted by the end time of the risk intervals. The columns for $\bar{x}$ and hazard are the values at the event times; events occur at the end of each interval for which status = 1.

4.1 Breslow approximation

For the Breslow approximation we have

$$\begin{aligned} LL &= \log\frac{r}{r+1} + \log\frac{1}{r+2} + \log\frac{1}{3r+2} + \\ &\quad \log\frac{r}{3r+1} + \log\frac{1}{3r+1} + 2\log\frac{r}{3r+2} \\ &= 4\log(r+1) - \log(r+3) - 3\log(3r+2) - 2\log(3r+1). \end{aligned}$$

$$\begin{aligned} U &= 1\left(\frac{r}{r+1} + 0\left(\frac{r}{r+2} + 0\left(\frac{3r}{3r+2} + \right.\right.\right. \\ &\quad \left.\left.\left. 1\left(\frac{3r}{3r+1} + 0\left(\frac{3r}{3r+1} + 2\left(1\left(\frac{3r}{3r+2}\right.\right.\right.\right.\right. \end{aligned}$$

$$\begin{aligned} H &= \frac{r}{(r+1)^2} + \frac{2r}{(r+2)^2} + \frac{6r}{(3r+2)^2} + \frac{3r}{(3r+1)^2} \\ &\quad \frac{3r}{(3r+1)^2} + \frac{12r}{(3r+2)^2}. \end{aligned}$$

In this case U is a quartic equation and we find the solution numerically.

```
> ufun <- function(r) {
      4 - (r/(r+1) + r/(r+2) + 3*r/(3*r+2) + 6*r/(3*r+1) + 6*r/(3*r+2))
  }
> rhat <- uniroot(ufun, c(.5, 1.5), tol=1e-8)$root
> bhat <- log(rhat)
> c(rhat=rhat, bhat=bhat)
       rhat        bhat
 0.91894769 -0.08452608
```

The solution is at $U(\hat{\ }) = 0$ or r .9189477; $\hat{\ } = \log(r)$ .084526. Then

$$\begin{aligned} LL(0) &= 9.392662 & LL(\hat{\ }) &= 9.387015 \\ U(0) &= 2=15 & U(\hat{\ }) &= 0 \\ H(0) &= 2821=1800 & H(\hat{\ }) &= 1.586934 \end{aligned}$$

The martingale residuals are (status{cumulative hazard) or $O - E = \sum_i \int Y_i(s) r_i d\hat{\ }(s)$. Let $\hat{\ }_1, \ldots, \hat{\ }_6$ be the six increments to the cumulative hazard listed in Table 2. Then the cumulative hazards and martingale residuals for the subjects are as follows.

Subject	$_i$	$M(0)$	$M(\hat{})$
1	$r\hat{}_1$	1{30/60	0.521119
2	$\hat{}_2$	1{20/60	0.657411
3	$\hat{}_3$	1{12/60	0.789777
4	$r(\hat{}_2 + \hat{}_3 + \hat{}_4)$	1{47/60	0.247388
5	$\hat{}_1 + \hat{}_2 + \hat{}_3 + \hat{}_4 + \hat{}_5$	1{92/60	-0.606293
6	$r\ \ (\hat{}_5 + \hat{}_6)$	1{39/60	0.369025
7	$r\ \ (\hat{}_3 + \hat{}_4 + \hat{}_5 + \hat{}_6)$	1{66/60	-0.068766
8	$r\ \ (\hat{}_3 + \hat{}_4 + \hat{}_5 + \hat{}_6)$	0{66/60	-1.068766
9	$\hat{}_6$	0{24/60	-0.420447
10	$\hat{}_6$	0{24/60	-0.420447

The score and Schoenfeld residuals can be laid out in a tabular fashion. Each entry in the table is the value of $fx_i \quad \bar{x}(t_j)gdM_i(t_j)$ for subject i and event time t_j. The row sums of the table are the score residuals for the subject; the column sums are the Schoenfeld residuals at each event time. Below is the table for $\quad = \log(2)$ ($r = 2$). This is a slightly more stringent test than the table for $\quad = 0$, since in this latter case a program could be missing a factor of $r = \exp(\ ) = 1$ and give the correct answer. However, the results are much more compact than those for $\hat{}$, since the solutions are exact fractions.

	Event Time						Score
Id	2	3	6	7	8	9	Resid
1	$\frac{1}{9}$						$\frac{1}{9}$
2		$\frac{3}{8}$					$\frac{3}{8}$
3			$\frac{21}{32}$				$\frac{21}{32}$
4		$\frac{1}{4}$	$\frac{1}{16}$	$\frac{5}{49}$			$\frac{165}{784}$
5	$\frac{2}{9}$	$\frac{1}{8}$	$\frac{3}{32}$	$\frac{6}{49}$	$\frac{36}{49}$		$\frac{2417}{14112}$
6					$\frac{2}{49}$	$\frac{1}{8}$	$\frac{33}{392}$
7			$\frac{1}{16}$	$\frac{2}{49}$	$\frac{2}{49}$	$\frac{1}{8}$	$\frac{15}{784}$
8			$\frac{1}{16}$	$\frac{2}{49}$	$\frac{2}{49}$	$\frac{1}{8}$	$\frac{211}{784}$
9						$\frac{3}{16}$	$\frac{3}{16}$
10						$\frac{3}{16}$	$\frac{3}{16}$
	$\frac{1}{3}$	$\frac{1}{2}$	$\frac{3}{4}$	$\frac{1}{7}$	$\frac{6}{7}$	$\frac{1}{2}$	$\frac{95}{84}$
	$\frac{1}{r+1}$	$\frac{-r}{r+2}$	$\frac{-3r}{r+2}$	$\frac{1}{3r+1}$	$\frac{3r}{3r+1}$	$\frac{4}{3r+2}$	

Both the Schoenfeld and score residuals sum to the score statistic $U(\ )$. As discussed further above, programs will return two Schoenfeld residuals at time 7, one for each subject who had an event at that time.

Time	Status	X	Wt	$\bar{x}(t)$	$\hat{d}_0(t)$
1	1	2	1	$(2r^2 + 11r)\hat{d}_0 = \bar{x}_1$	$1=(r^2 + 11r + 7)$
1	0	0	2		
2	1	1	3	$11r=(11r + 5) = \bar{x}_2$	$10=(11r + 5)$
2	1	1	4		
2	1	0	3		
2	0	1	2		
3	0	0	1		
4	1	1	2	$2r=(2r + 1) = \bar{x}_3$	$2=(2r + 1)$
5	0	0	1		

Table 3: Test data 3

4.2 Efron approximation

This example has only one tied death time, so only the term(s) for the event at time 9 change. The main quantities at that time point are as follows.

	Breslow	Efron
LL	$2\log\frac{r}{3r+2}$	$\log\frac{r}{3r+2} + \log\frac{r}{2r+2}$
U	$\frac{2}{3r+2}$	$\frac{1}{3r+2} + \frac{1}{2r+2}$
H	$2\frac{6r}{(3r+2)^2}$	$\frac{6r}{(3r+2)^2} + \frac{4r}{(2r+2)^2}$
$\hat{d}$	$\frac{2}{3r+2}$	$\frac{1}{3r+2} + \frac{1}{2r+2}$

5 Test data 3

This is very similar to test data 1, but with the addition of case weights. There are 9 observations, x is a 0/1/2 covariate, and weights range from 1 to 4. As before, let $r = \exp(\ )$ be the risk score for a subject with $x = 1$. Table 3 shows the data set along with the mean and increment to the hazard at each point.

5.1 Breslow estimates

The likelihood is a product of terms, one for each death, of the form

$$\left(\frac{e^{X_i}}{\sum_j Y_j(t_i) w_j e^{X_j}}\right)^{w_i}$$

For integer weights, this gives the same results as would be obtained by replicating each observation the specified number of times, which is in fact one motivation for the definition. The definitions for the score vector U and information matrix H simply replace the mean and variance with weighted versions of the same. Let $PL(\ ; w)$ be the log partial liklihood when all the observations are given a common case weight of w; it is easy to prove that $PL(\ ; w) = wPL(\ ; 1) \quad d\log(w)$

where d is the number of events. One consequence of this is that PL can be positive for weights that are less than 1, a case which sometimes occurs in survey sampling applications. (This can be a big surprise the first time one encounters it.)

$$\begin{aligned} LL &= f2 \quad \log(r^2 + 11r + 7)g + 3f \quad \log(11r + 5)g \\ &\quad +4f \quad \log(11r + 5)g + 3f0 \quad \log(11r + 5)g \\ &\quad +2f \quad \log(2r + 1)g \\ &= 11 \quad \log(r^2 + 11r + 7) \quad 10\log(11r + 5) \quad 2\log(2r + 1) \end{aligned}$$

$$\begin{aligned} U &= (2 \quad \bar{x}_1) + 3(0 \quad \bar{x}_2) + 4(1 \quad \bar{x}_2) + 3(1 \quad \bar{x}_2) + 2(1 \quad \bar{x}_3) \\ &= 11 \quad [(2r^2 + 11r)=(r^2 + 11r + 7) + 10(11r=(11r + 5)) + 2(2r=(2r + 1))] \\ I &= [(4r^2 + 11r)=(r^2 + 11r + 7) \quad \bar{x}_1^2] + 10(\bar{x}_2 \quad \bar{x}_2^2) + 2(\bar{x}_3 \quad \bar{x}_3^2) \end{aligned}$$

The solution corresponds to $U(\ ) = 0$ and can be computed using a simple search for the zero of the equation.

```
> ufun <- function(r) {
      xbar <- c( (2*r^2 + 11*r)/(r^2 + 11*r +7), 11*r/(11*r + 5), 2*r/(2*r +1))
      11- (xbar[1] + 10* xbar[2] + 2* xbar[3])
  }
> rhat <- uniroot(ufun, c(1,3), tol= 1e-9)$root
> bhat <- log(rhat)
> c(rhat=rhat, bhat=bhat)
     rhat      bhat
2.3621151 0.8595574
```

From this we have

```
> wfun <- function(r) {
      beta <- log(r)
      pl <- 11*beta - (log(r^2 + 11*r + 7) + 10*log(11*r +5) + 2*log(2*r +1))
      xbar <- c((2*r^2 + 11*r)/(r^2 + 11*r +7), 11*r/(11*r +5), 2*r/(2*r +1))
      U <- 11 - (xbar[1] + 10*xbar[2] + 2*xbar[3])
      H <- ((4*r^2 + 11*r)/(r^2 + 11*r +7)- xbar[1]^2) +
          10*(xbar[2] - xbar[2]^2) +  2*(xbar[3]- xbar[3]^2)
      c(loglik=pl, U=U, H=H)
  }
> temp <- matrix(c(wfun(1), wfun(rhat)), ncol=2,
         dimnames=list(c("loglik", "U", "H"), c("beta=0", "beta-hat")))
> round(temp, 6)
           beta=0   beta-hat
loglik -32.867551 -32.021046
U        2.107456   0.000000
H        2.914212   1.966555
```

When = 0, the three unique values for $\bar{x}$ at t = 1, 2, and 4 are 13/19, 11/16 and 2/3, respectively, and the increments to the cumulative hazard are 1/19, 10/16 = 5/8, and 2/3, see table 3. The martingale and score residuals at = 0 and ^ are

Id	Time	$M(0)$	$M(\hat{})$
A	1	1 1=19 = 18=19	0.85531
B	1	0 1=19 = 1=19	-0.02593
C	2	1 (1=19 + 5=8) = 49=152	0.17636
D	2	1 (1=19 + 5=8) = 49=152	0.17636
E	2	1 (1=19 + 5=8) = 49=152	0.65131
F	2	0 (1=19 + 5=8) = 103=152	-0.82364
G	3	0 (1=19 + 5=8) = 103=152	-0.34869
H	4	1 (1=19 + 5=8 + 2=3) = 157=456	-0.64894
I	5	0 (1=19 + 5=8 + 2=3) = 613=456	-0.69808

Score residuals at = 0 are

Id	Time	Score
A	1	(2 13=19)(1 1=19)
B	1	(0 13=19)(0 1=19)
C	2	(1 13=19)(0 1=19) + (1 11=16)(1 5=8)
D	2	(1 13=19)(0 1=19) + (1 11=16)(1 5=8)
E	2	(0 13=19)(0 1=19) + (0 11=16)(1 5=8)
F	2	(1 13=19)(0 1=19) + (1 11=16)(0 5=8)
G	3	(1 13=19)(0 1=19) + (0 11=16)(0 5=8)
H	4	(1 13=19)(0 1=19) + (1 11=16)(0 5=8)
		+(1 2=3)(1 2=3)
I	5	(1 13=19)(0 1=19) + (1 11=16)(0 5=8)
		+(0 2=3)(0 2=3)

R also returns unweighted residuals by default, with an option to return the weighted version; it is the weighted sum of residuals that totals zero, $\sum w_i M_i = 0$. Whether the weighted or the unweighted form is more useful depends on the intended application, neither is more \correct" than the other. R does di er for the dfbeta residuals, for which the default is to return weighted values. For the third observation in this data set, for instance, the unweighted dfbeta is an approximation to the change in ^ that will occur if the case weight is changed from 2 to 3, corresponding to deletion of one of the three \subjects" that this observation represents, and the weighted form approximates a change in the case weight from 0 to 3, i.e., deletion of the entire observation.

The increments of the Nelson-Aalen estimate of the hazard are shown in the rightmost column of table 3. The hazard estimate for a hypothetical subject with covariate $X^\dagger$ is $_i(t) = \exp(X^\dagger\ )\ _0(t)$ and the survival estimate is $S_i(t) = \exp(\ _i(t))$. The two term of the variance, for $X^\dagger = 0$, are Term1 + $d'Vd$:

Time	Term 1
1	$1=(r^2 + 11r + 7)^2$
2	$1=(r^2 + 11r + 7)^2 + 10=(11r + 5)^2$
4	$1=(r^2 + 11r + 7)^2 + 10=(11r + 5)^2 + 2=(2r + 1)^2$

Time	d
1	$(2r^2 + 11r)=(r^2 + 11r + 7)^2$
2	$(2r^2 + 11r)=(r^2 + 11r + 7)^2 + 110r=(11r + 5)^2$
4	$(2r^2 + 11r)=(r^2 + 11r + 7)^2 + 110r=(11r + 5)^2 + 4r=(2r + 1)^2$

For = log(2) and $X^{\dagger} = 0$, where k the variance of ^ = 1/2.153895 this reduces to

Time	Variance	
1	1/1089	$+ k(30=1089)^2$
2	(1/1089+ 10/729)	$+ k(30=1089 + 220=729)^2$
4	(1/1089+ 10/729 + 2/25)	$+ k(30=1089 + 220=729 + 8=25)^2$

giving numeric values of 0.0012706, 0.0649885, and 0.2903805, respectively.

5.2 Efron approximation

For the Efron approximation the combination of tied times and case weights can be approached in at least two ways. One is to treat the case weights as replication counts. There are then 10 tied deaths at time 2 in the data above, and the Efron approximation involves 10 di erent denominator terms. Let $a = 7r + 3$, the sum of risk scores for the 3 observations with an event at time 2 and $b = 4r + 2$, the sum of risk scores for the other subjects at risk at time 2. For the replication approach, the loglikelihood is

$$\begin{aligned} LL = & f2 \quad \log(r^2 + 11r + 7)g + \\ & f7 \quad \log(a + b) \quad \log(:9a + b) \quad ::: \quad \log(:1a + b)g + \\ & f2 \quad \log(2r + 1) \quad \log(r + 1)g: \end{aligned}$$

A test program can be created by comparing results from the weighted data set (9 observations) to the unweighted replicated data set (19 observations). This is the approach taken by SAS `phreg` using the `freq` statement. It's advantage is that the appropriate result for all of the weighted computations is perfectly clear the disadvantage is that the only integer case weights are supported. (A secondary advantage is that I did not need to create another algebraic derivation for this appendix.)

A second approach, used in the survival package, allows for non-integer weights. The data is considered to be 3 tied observations, and the log-likelihood at time 2 is the sum of 3 weighted terms. The rst term of the three is one of

$$\begin{aligned} & 3[\quad \log(a + b)] \\ & 4[\quad \log(a + b)] \\ \text{or} \quad & 3[0 \quad \log(a + b)]; \end{aligned}$$

depending on whether the event for observation C, D or E actually happened first (had we observed the time scale more exactly); the leading multiplier of 3, 4 or 3 is the case weight. The second term is one of

$$\begin{array}{rl} & 4[\quad \log(4s + 3 + b)] \\ & 3[0 \quad \log(4s + 3 + b)] \\ & 3[\quad \log(3s + 3 + b)] \\ & 3[\quad \log(3s + 3 + b)] \\ & 3[0 \quad \log(4s + 3 + b)] \\ \text{or} \quad & 4[\quad \log(4s + 3 + b)]. \end{array}$$

The first choice corresponds to an event order of observation C then D (subject D has the event, with D and E still at risk), the second to $C \; ! \; E$, then $D \; ! \; C$, $D \; ! \; E$, $E \; ! \; C$ and $E \; ! \; D$, respectively. For a weighted Efron approximation first replace the argument to the log function by its average argument, just as in the unweighted case. Once this is done the average term in the above corresponds to using an average weight of 10/3.

The final log-likelihood and score statistic are

$$\begin{aligned} LL &= f2 \quad \log(r^2 + 11r + 7)g \\ &\quad + f7 \quad (10=3)[\log(a + b) + \log(2a=3 + b) + \log(a=3 + b)]g \\ &\quad + 2f \quad \log(2r + 1)g \end{aligned}$$

$$\begin{aligned} U &= (2 \quad \bar{x}_1) + 2(1 \quad \bar{x}_3) \\ &\quad + 7 \quad (10=3)[\bar{x}_2 + 26r=(26r + 12) + 19r=(19r + 9)] \\ &= 11 \quad (\bar{x}_1 + (10=3)(\bar{x}_2 + \bar{x}_{2b} + \bar{x}_{2c}) + 2\bar{x}_3) \end{aligned}$$

$$\begin{aligned} I &= [(4s^2 + 11s)=(s^2 + 11s + 7) \quad \bar{x}_1^2] \\ &\quad + (10=3)[(\bar{x}_2 \quad \bar{x}_2^2) + (\bar{x}_{2b} \quad \bar{x}_{2b}^2) + (\bar{x}_{2b} \quad \bar{x}_{2b}^2) \\ &\quad + 2(\bar{x}_3 \quad \bar{x}_3^2) \end{aligned}$$

The solution is at $= .87260425$, and

$$\begin{array}{ll} LL(0) = \quad 30.29218 & LL(\hat{}) = \quad 29.41678 \\ U(0) = 2.148183 & U(\hat{}) = 0 \\ H(0) = 2.929182 & H(\hat{}) = 1.969447. \end{array}$$

The hazard increment and mean at times 1 and 4 are identical to those for the Breslow approximation, as shown in table 3. At time 2, the number at risk for the first, second and third portions of the hazard increment are $n_1 = 11r + 5$, $n_2 = (2=3)(7r + 3) + 4r + 2 = (26r + 12)=3$, and $n_3 = (1=3)(7r + 3) + 4r + 2 = (19r + 9)=3$. Subjects F{I experience the full hazard at time 2 of $(10=3)(1=n_1 + 1=n_2 + 1=n_3)$, subjects B{D experience $(10=3)(1=n_1 + 2=3n_2 + 1=3n_3)$. Thus, at $= 0$ the martingale residuals are

Id	Time	$M(0)$	
A	1	1 - 1/19	= 18/19
B	1	0 - 1/19	= -1/19
C	2	1 - (1/19 + 10/48 + 20/114 + 10/84)	=473/1064
D	2	1 - (1/19 + 10/48 + 20/114 + 10/84)	=473/1064
E	2	1 - (1/19 + 10/48 + 20/114 + 10/84)	=473/1064
F	2	0 - (1/19 + 10/48 + 10/38 + 10/28)	=-2813/3192
G	3	0 - (1/19 + 10/48 + 10/38 + 10/28)	=-2813/3192
H	4	1 - (1/19 + 10/48 + 10/38 + 10/28 + 2/3)	=-1749/3192
I	5	0 - (1/19 + 10/48 + 10/38 + 10/28 + 2/3)	=-4941/3192

The hazard estimate for a hypothetical subject with covariate $X^\dagger$ is $_i(t) = \exp(X^\dagger\ )\ _0(t)$, $_0$ has increments of $1=(r^2 + 11r + 7$, $(10=3)(1=n_1 + 1=n_2 + 1=n_3)$ and $2=(2r + 1)$. This increment at time 2 is a little larger than the Breslow jump of $10=d1$. The first term of the variance will have an increment of $[\exp((X^\dagger\ )(]^2(10=3)(1=n_1^2 + 1=n_2^2 + 1=n_3^2)$ at time 2. The increment to the cumulative distance from the center d will be

$$
\begin{aligned}
& [X^\dagger \quad \frac{11r}{11r + 5}]\frac{10}{3n_1} \\
+ \quad & [X^\dagger \quad \frac{(2=3)7r + 4r}{n2}](10=3)(1=n_2) \\
+ \quad & [X^\dagger \quad \frac{(1=3)7r + 4r}{n2}](10=3)(1=n_3)
\end{aligned}
$$

For $X^\dagger = 1$ and $\ = \ =3$ we get cumulative hazard and variance below. We have $r \quad e \ =3$, $V =$

6 Multi-state data

Figure 1 shows a simple multi-state model, while table 4 shows a data set for the model. Subject 1 follows the path of Entry, a, b, a, with no further follow up after the final transition, while subjects 4 and 5 are 'censored'; they have further follow-up after the last observed change of state.

References

[1] B. Efron. The efficiency of Cox's likelihood function for censored data. *J. Amer. Stat. Assoc.*, 72:557{565, 1977.

[2] T. R. Fleming and D. P. Harrington. Nonparametric estimation of the survival distribution in censored data. *Comm. Stat. Theory Methods*, 13:2469{2486, 1984.

[3] T. M. Therneau and P. M. Grambsch. *Modeling Survival Data: Extending the Cox Model*. Springer-Verlag, New York, 2000.

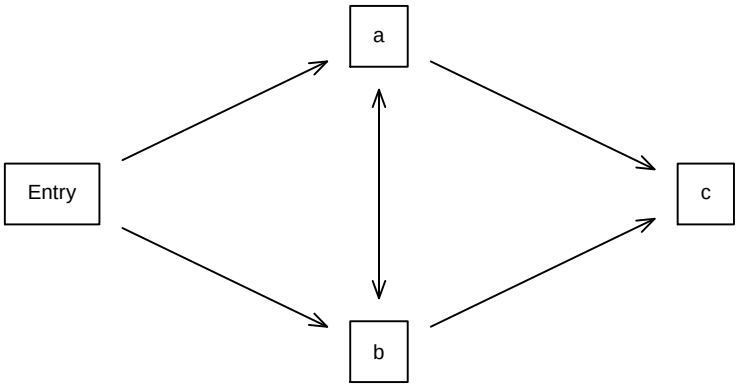

Figure 1: A simple 4 state model.

id	time1	time2	event	x
1	0	4	a	0
1	4	9	b	0
1	9	10	a	0
2	0	5	b	1
3	2	9	c	1
4	0	2	a	0
4	2	8	c	0
4	8	9		0
5	1	3	b	2
5	3	11		2

Table 4: A multi-state data set.