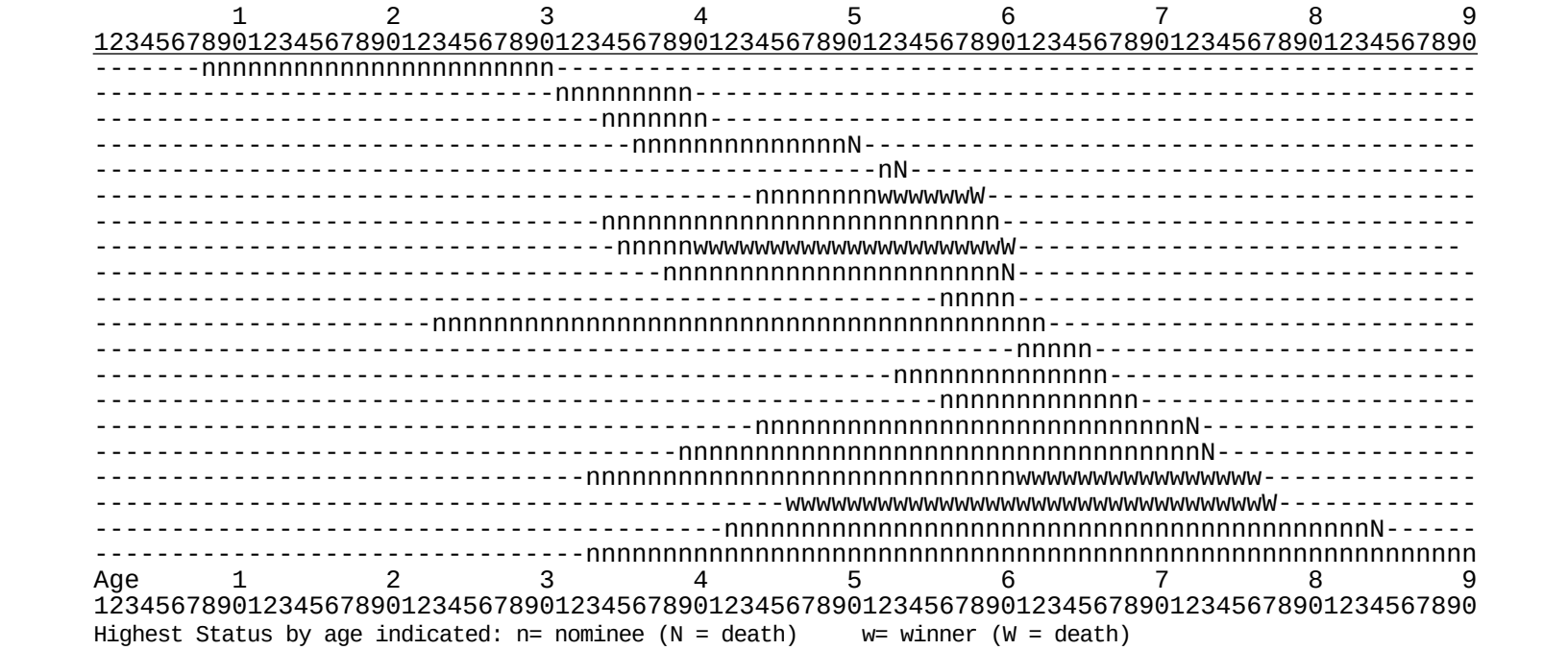
Course EPIB-634: Survival Analysis & Related Topics [Winter 2007] Assignment 8

[See SAS/Stata/R programs provided under Resources in www.epi.mcgill.ca/hanley/c681/cox

1 We revisit some of the 'Oscars' data used in Assignment 4. The year of birth, age at first Oscar nomination, age at first win, age in 2001 or at death, and vital status, for 20 of the male performers were as follows

IDENTITY	YEAR.BIRTH	AGE_NOM1	AGE_WIN1	AGE_LAST	DEAD
859	1971	8	.	30	0
1290	1962	31	.	39	0
1502	1961	34	.	40	0
1457	1930	36	.	50	1
690	1888	52	.	53	1
628	1899	44	52	58	1
535	1942	34	.	59	0
493	1901	35	40	60	1
786	1929	38	.	60	1
1340	1941	56	.	60	0
1487	1939	23	.	62	0
177	1936	61	.	65	0
308	1935	53	.	66	0
291	1933	56	.	68	0
1525	1916	44	.	72	1
1009	1923	39	.	73	1
1231	1925	33	61	76	0
1052	1919	46	46	77	1
1082	1888	42	.	84	1
627	1911	33	.	90	0

These 20 life-courses are re-arranged below, with the time (age) axis running from left to right.



Earlier, you calculated a summary ratio comparing the mortality experience in the person-time lived as nominees and in the person time lived as winners i.e. with winner as a time-varying covariate. The task now is to obtain a hazard ratio using the proportional hazards model with "winner" as a time-varying covariate, using for a-c the data on these 20 selected performers, and for d-f the entire dataset.

a Set up, in R or Stata or SAS, (possibly split) records*, each containing the 3 t.start, t.stop and event variates, as well as the time-dependent variate "winner" and obtain the ML estimate of the HR using coxph / stcox /phreg - use Efron option for ties. (*modify the data vectors given on website)

b Which records/partial records could be omitted without changing the Likelihood?

c Add a vector (variate) to denote the year of birth of the 'owner' of each record, and refit using year of birth as a fixed variate, and winner as a time-dependent variate.

d Fit a "winner time-dependent, male, year.born" ph model to the *entire* set of winners & nominated. (cf 2nd full para. of 2nd column of page W98 of the Appendix (used for "New Analysis PH" row in table on p363 of article) [dataset in Resources] Comment. Interpret also the exponentiated coefficients of the other two variates in the model (male and year.born)

e Repeat d, but using a "stratified Cox mode -- ie use male/female as strata, year.born as fixed, and winner as time-dependent.

f How different are the HR estimates for "winner" in d and e from that obtained using a ph model with the same variates, but with winner treated as a fixed variate? Explain the *direction* of the difference.

2 Refer to the article "Weekend versus Weekday Admission and Mortality from Myocardial Infarction". Restrict attention to the data for 1999-2002.

In Assignment 7, you calculated day-specific hazards (or incidence density) for each of the first seven days, and examined the day-by-day (crude) hazard ratios . This dataset had so many deaths (several hundred per day) that the HR pattern was clear from the raw data. But what if (in a much smaller study) there were fewer deaths? Could we model the TREND in the HR's over days (time) as a way to see how far from constant they are?

One way to do this is to divide up the time scale (again using split records), so that we can use (in some kind of regression model) the time to which each HR refers. In this case, it is *not* the "x" *variable* (here admitted on weekday vs. weekend) that is changing (time dependent) but rather the value of the *beta coefficient* associated with the "x" We can say that "t" *modifies* beta, and model non-proportional hazards (non-constant hazard ratios) using product terms involving t (as we do with product terms in conventional regression models).

Even though it is clear from the data that the HR's follow a step-function shape (1.2 or so for 4 days, then approx. 1 thereafter) that is not easily represented by a smooth curve, we can use the data to practice fitting smooth curves for non-proportional hazards (non-constant hazard ratios (the fits will not be good but..)

i Follow the new R code in the website to first create a 'split records' dataset suitable for fitting ph (and non-p h) models. Make sure you list the dataset *for yourself* and understand how we can use a "weight" in the dataset that tells us how many persons each record represents (helpful when there are lots of persons -- the same facility is available in SAS and Stata)

Inspect (and understand the purpose of) the `t.weekend.product` and `t.sq.weekend.product` terms.

ii Fit the constant HR, linear log(HR), and curvilinear log(HR) models using the models provided. Comment.

iii While JH was setting up these models, he fitted one other one. After defining

```
ds$day = ds$t.start-1;
```

he asked for

```
fit.unfittablemodel = coxph(Surv(t.start,t.stop,dead) ~ weekend + day,
weights=freq, data=ds);
```

Can you see why it cannot be fitted? Hint; all comparisons are within-riskset comparisons! The estimation method uses the riskset for each unique event time (here a day) as a separate stratum.

3 Do the 'risk factor' levels measured in 1948 in the Framingham Heart Study lose prognostic ability with (follow-up) time?

The Framingham dataset on the web covers the first 20+ years from 1948, and so we might expect the regression *coefficients* associated with the baseline measurements to 'degade' with follow-up time, i.e. we expect a TREND in the (log) HR's over the years? [some measurements were repeated at each followup, and raise a *separate* question as to how these time-dependent *covariates* should be used. They are not in this dataset]

i We would expect the HR's to follow a smooth curve, so you are asked to fit a linear trendline in the *cholesterol* beta over time. Note the use of centering (via a z score) to avoid large product terms. *Note also that we cannot have a main effect of time here*, since *we are not modeling how the hazard rate varies with time*, only how the hazard *ratio* varies with time. We have repeatedly seen that the Cox model uses a series of time-*matched* risksets, just like strata are used in the MH approach. However, since we don't specify the hazard function for the covariate vector x=0, we cannot study the main effect of time (If we fitted a fully parametric rate model, with terms for time, we could!)

ii Select what you expect will be two other strong prognostic factors, and assess whether there is evidence that their coefficients in the hazard model change with time. Comment on your findings.