Expected Survival Based on Hazard Rates (Update)

Terry Therneau Jan Offord

February 18, 1999

Technical Report Number 63
Section of Biostatistics
Mayo Clinic

1 Introduction

This paper is an extension and update of Technical Report #52 [13]. An update to the rate tables themselves is based on the recently released data from the 1990 decennial census [18], which allowed us to replace extrapolated 1990 death rates with actual rates, and to improve the extrapolated year 2000 values. Much of the material in the prior report is contained here, in order to make this document useful on it's own.

The expected survival computations are based on a set of tables containing survival probabilities for the US population. These tables have been compiled over several years by members of the Department of Health Sciences Research; earlier versions are documented in Bergstralh and Offord [2], a SAS procedure that makes use of them was a part of the SAS Supplemental Library [12]. (These procedures are no longer distributed by SAS, however). Details of these data sets are discussed in section 2.

Sections 3 and 4 of the report gives background on the computation of an expected survival curve based on these data sets. There are several methods, each with its advantages. The methods and their relative merits seem to be "rediscovered" on a regular basis in the literature. Sections 5 and 6 discuss S-Plus and SAS functions that implement these techniques. Examples are given that use both the US population and user-created rate tables.

2 Expected Survival Rates

The expected survival data consists of 5 groups of tables: US, Minnesota, Florida, Arizona, and West North Central (WNC). The WNC region consists of the states Nebraska, Kansas, Missouri, North and South Dakota, Iowa and Minnesota. All are divided by age, sex and calendar year, with optional further divisions by race, and are derived from published US and regional mortality data. The data tables are published for decade years, usually with about a 5-7 year lag, e.g., we expect to have the year 2000 data available by 2006. Each table is based on the average of 3 years, e.g., 1989-1991. The table entry $q_{1960,24,\mathrm{F}}$ would contain the probability that a female who became 24 years old sometime in 1960 will die on or before her 25th birthday. The value of .16859 for age 84–45 white males in 1940 states that approximately 17% of the men who became 84 years old sometime during 1940 perished before reaching their 85th birthday.

2.1 United States

The S-Plus data sets are `survexp.us` and `survexp.usr`. The first is a 3 way array with dimensions of age (0–1 day, 1–7 days, 7–28 days, 28 days–1 year, 1–2 years, ... 109–110 years), gender ('male', 'female') and calendar year (1940, 1950, ..., 2000). The `survexp.usr` table has dimensions of age, sex, race ('white', 'nonwhite', 'black') and year. The year 2000 data is an extrapolation, which is discussed below. The sources for the tables are

1950 United States Life Tables and Actuarial Tables 1939-1941, Federal Security Agency, United States Public Health Service, National Office of Vital Statistics, US Government Printing Office, 1947.

1950 Life Tables for 1949-51, Vital Statistics, Special Reports, US Department of Health, Education, and Welfare, Public Health Service, Volume 41, No 1, 1956.

1960 United States Lifetables 1959-61, Public Health Service Publication No. 1252, Volume 1, No. 1

1970 U.S. Decennial Lifetables 1969-71, DHEW Publication No. HRA 75-1150, Volume 1, No. 1

1980 U.S. Decennial Lifetables 1979-81, DHEW Publication No. PHS 85-1150-1, Volume 1, No. 1

1990 National Center for Health Statistics, U.S. Decennial Lifetables 1989-91, Volume 1, No. 1, Hyattsville, Maryland, 1997.

At the time of this writing, the more recent tables could be found at the National Center for Health Statistics web site, http://www.cdc.gov/nchswww, by following the links for products → published reports → life tables.

The prior version of these tables had both a `survexp.uswhite` and `survexp.usr` data set, the former having years 1950–2000 and the latter years 1960–2000. The new `survexp.usr` data set subsumes both of these.

There have been some changes in definitions over the time period covered by the tables, and creation of a single table involves some compromises. The breakdown by race has been

- 1940: white, negro, other than white or negro
- 1950–60: white, non-white
- 1970–90: white, non-white, black

Based on comparisons of the the three groups in 1940, 70, 80 and 90, the 'black' dimension for 1950 and 1960 is a duplicate of the non-white data.

The 1940 data breaks down the first year of life into multiple intervals: daily ending on 1, 2, 3, 7, 14, 21, and 28 days, and monthly ending at 1–12 months. The probabilty of dying in days 1–7 is then $1-\{(1-q_1)(1-q_3)(1-q_7)\}$, with similar computations for the other intervals. The table for white males only goes to age 108–109 (one entry short). The tables for non-white males and females both end at the 105–106 interval. Because the probability of death q is $> 70\%$ in both the male and female tables, implying very few survivors to or past this point, the table was filled out by propagating the age 105 value forward. Interestingly, the table for black females (but none other) extends to age 112-113; only the values up to age 109–110 are in the S-Plus data set. (A linear regression on $\log(q)$ for the last few years can be extrapolated forward to give values for these later ages, but gives values of $q > 1$ for some of the points).

The 1960 decennial data has a different set of intervals for the first year of life: 0-1, 1-3, 3-28, and 28 days - 1 year. However, a different source [19], based on only the 1960 rather than 1959–61 data, contains the necessary breakdown. The actual values, for day 0–1 for instance, differ slightly from the decennial data; we used the new source as a relative scale for interpolating the missing interval in the decennial table.

The 1970 data sets and onwards use the same breakdown for the first year of life as the S-Plus tables.

The SAS data set is `lt_us`, and contains variables age (0–109), sex('f', 'm'), year (1950–2000), and race ('t', 'b', 'nw', 'w'). The race codes are "total" (all races), "black", "non-white", and "white". The SAS data set does not subdivide the first year of life. It contains only total and white for 1950, all but black for 1960 and 1970, and all four categories thereafter.

2.1.1 West North Central

The WNC data set `survexp.wnc` has dimensions of age (0–.5, .5–1, 1–109), sex, and decade calendar years 1910–2000, and contains rates for the white population of the region. The SAS data set `lt_wnc` does not subdivide the first year of life.

In the years prior to 1950 a separate Minnesota table was not issued, presumably because the denominator population was too small, particularly in the older age groups. From 1970 onward a WNC table has not been published, and our WNC table contains Minnesota white data. The main use of this table is in conjunction with long-term studies associated with the Rochester Epidemiology Project. (A recent study of hip fracture, for instance, included all incident cases from 1928 to 1982 inclusive and examined changes in post-fracture survival). The table may be of less interest outside of the institution. See table 1 in [2] for details on the sources and computations used for the earlier years.

The 1990 Minnesota data does not include information to subdivide the first year of life. Data for this was taken from the infant mortality data, 1989-91 average, table 2-14 of [16], which shows for each of the 50 states the proportion of first year deaths which occurred in 0-6 months (77/100 and 44/55 for males and females, respectively). This proportion was used to divide the first year's hazard.

2.2 Minnesota

The 1990 Minnesota life tables are the first to include separate data for non-whites, thus the S-Plus data includes tables only for the total population (`survexp.mn`) and for the total white population (`survexp.mnwhite`). The first of these is given only for 1970–2000, and is thus equal to the West North Central table for all but the first year of life, where the WNC table is subdivided into the first and second half years. The Minnesota white table is given for 1950–2000.

1950 Minnesota State Life Tables 1949-51, Vital Statistics, Special Reports, US Department of Health, Education, and Welfare, Public Health Service, Volume 41, Supplement 22, 1956.

1960 Minnesota State Life Tables 1959-61, Public Health Service Publication No. 1252, Volume 2, No. 24

1970 U.S. Decennial Lifetables 1969-71, DHEW Publication No. HRA 75-1151, Volume 2, No. 24

1980 U.S. Decennial Lifetables 1979-81, DHEW Publication No. PHS 86-1151-24, Volume 2, No. 24

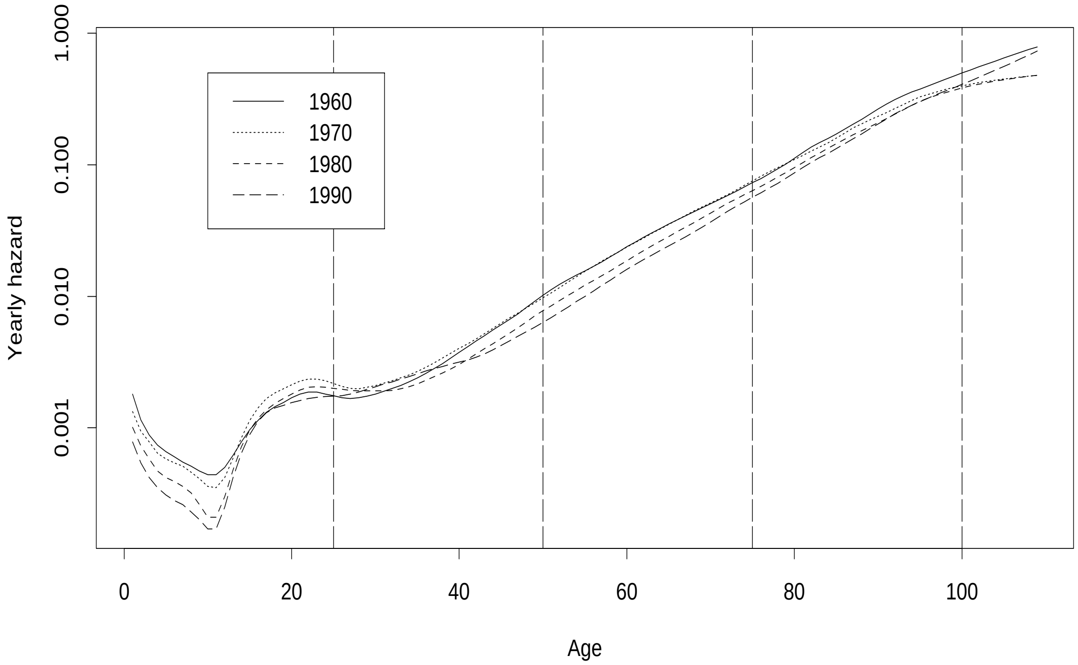

Figure 1: *Hazard rates for U.S. Males*

1990 National Center for Health Statistics, U.S. Decennial Lifetables 1989-91, Volume II, State life tables no. 24, Minnesota, Hyattsville, Maryland, 1998.

The SAS data set is `lt_mn`, with variables age (0–109), sex ('f', 'm'), year and race. There is data on `race='w'` for 1950–2000, for total population from 1970-2000, and non-white for 1990 and 2000.

2.3 Florida

We did not seek out any data sources before 1970 for the Florida data set, given Mayo's short presence there. The 1970 data has rates for total, white and non-white, the 1980 and 1990 publications added blacks. The S-Plus data sets are `survexp.fl` and `survexp.flr`, the former has dimensions of age, sex and year, the latter of age, sex, race and year. For the 1970 black dimension of the table, the 1970 non-white values were used. A plot shows the 1980 black data to be somewhat better approximated by 1980 non-white than by 1990 black, and we reasoned by analogy for 1970. The following S-Plus code draws the relevant curves.

```
> temp.male <- cbind(survexp.flr[,1,2:3,3], survexp.flr[,1,2,2])
> matplot(0:109, temp.male*365, type=’l’, log=’y’, col=1:3,
        xlab="Age", ylab="Yearly Hazard")
> legend(60, .01, c("1980 black", "1990 black", "1980 nonwhite"),
           lty=1:3, col=1:3)
> temp.female <- cbind(survexp.flr[,2,2:3,3], survexp.flr[,2,2,2])
> matplot(0:109, temp.female*365, type=’l’, log=’y’, col=1:3,
        xlab="Age", ylab="Yearly Hazard")
> legend(60, .01, c("1980 black", "1990 black", "1980 nonwhite"),
           lty=1:3, col=1:3)
```

The SAS data set `lt_fl` has variables of age, sex, year and race. There are 660 observations for 1970 (2 genders x 110 ages x 3 races) and 880 for the other 3 years. The data set does not fill in an “assumption” for the 1970 black population, but because of how the programs work, this is computationally equivalent to using the 1980 black data for 1970 black survival.

2.4 Arizona

The published 1990 data for Arizona contains all four race categories (total, white, non-white, black), the 1980 data contains the first 3, and the 1970 data contains only total and white. The S-Plus data set `survexp.az` contains total survival for 1970–2000, and the data set `survexp.azr` contains white and non-white for 1980–2000. The SAS data set `lt_az` contains all of the data.

2.5 Computer Tables

The S-Plus rate tables are contained in an object of class ‘ratetable’. This is essentially a multi-way array, with extra information included that allows the computing algorithms to distinguish between fixed margins, e.g. ‘sex’, which do not change over time, versus time-dependent margins such as ‘age’ and ‘year’ for which a subject changes categories over the course of his/her follow-up. All of the rate tables are by age, sex, calendar year, and optionally race, however, rate tables with other dimensions can be easily created. Only the decade calendar years are stored, data for intervening years is interpolated on demand.

To maintain backwards compatibility for old studies, the data sets `survexp.oldus`, `survexp.oldusr`, `survexp.oldwnc` and `survexp.oldmn` contain the prior versions of the data sets. Since the master files are maintained with SCCS, any of the old data sets could be retrieved on request if necessary.

The SAS data sets contain one observation per hazard value, and have the following variables:

`pop` = 3 character population name (US,MN,WNC,AZ,FL)

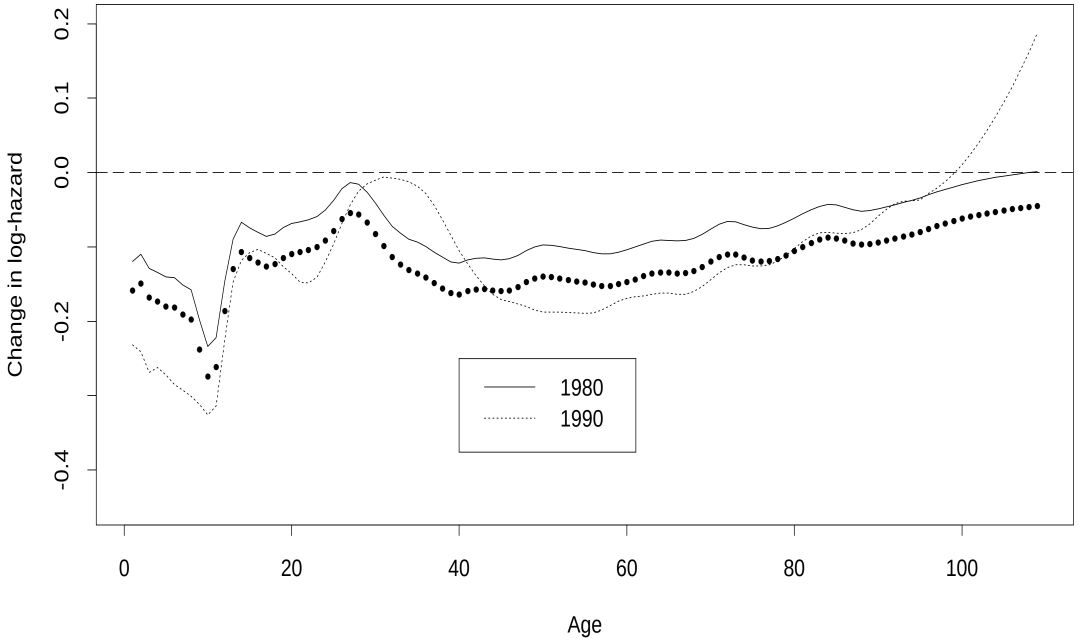

Figure 2: *Changes in log-hazard (base 10) between 1970 and 1990, US males*

`year` = decade specification (1910-2000)

`sex` = 1 character sex (m,f)

`race` = 2 character race (t=total, w=white, b=black, nw=non-white) Please note b and nw are not mutually exclusive.

`age` = age (0-109) (whole years only)

`q` = probability of dying before next birthday (from life table)

`hazard` = calculated daily hazard $= -\log(1-q)/365.241$

The SAS data sets are rarely accessed directly. The macros `%surv`, `%ltp` etc use `pop=us` for instance to reference the US population. A separate parameter `pop80=y` can be used to request the old set of rate tables.

3 Extrapolation

There is a time lag of 4-7 years between each census and the publication of the corresponding rate tables; we expect that the year 2000 tables will not become available until some time in 2006 or 2007. The expected survival functions use interpolation between calendar years within the rate table, but outside of the range of years they use the closest available date, e.g., if using the US total rate table then the expected number of events for a subject in 1910 would be assessed using the 1940 rates.

Given the continuing improvement in overall survival over the last 3 decades, use of the 1990 rates as a comparison for post-1990 data would be biased. Extrapolation of the risk of death, however, is perilous, as is any extrapolation of population data. The extrapolation method that we used for the year 2000 data was based therefore on two premises: to reduce the overall bias that would result from no extrapolation and to keep the model simple. Figure 1 shows the hazard rate as a function of time for United States males. Vertical lines have been drawn for reference purposes at ages 25, 50, 75 and 100.

The prior rate tables contained extrapolated values for both 1990 and 2000. The method used was aggressively simple: we noted that the hazards (as a function of age) for 1960, 70 and 80 were nearly evenly spaced on the log-hazard scale, with a mean difference of $.0979 + .00015 * \text{age}$ for males and $.1448 + .0005 * \text{age}$ for females. This correction, based on the US total data, was applied to all the 'total' rate tables in order to generate year 1990 and 2000 extrapolations. Similar corrections based on total white and Minnesota white were used for other rate tables. Details are in Therneau and Scheib [14].

The wide range of values makes differences between the years difficult to examine, so figure 2 plots the change in log hazard since 1970 as a function of age. For the males, the 1970 data shows substantial gains at ages 0–15 (the greatest at age 10) and 40–70, with moderate changes in survival for ages 20–35. The 1970 data is, by definition, the horizontal line. When the 1990 extrapolation (points) is compared to the actual 1990 data, we see that the extrapolation was a qualified success. The extrapolated values are closer to the actual values than the 1980 values were; without extrapolation the programs would use 1980 values by default. There were three areas of systematic error: although the predicted gain at age 80 is quite accurate, the predicted gains in survival beyond that age did not occur; the gains for infants age 0-5 were better than anticipated, and there was some increase in mortality for ages 28-40 in the males (AIDS?), with a lesser increase for females.

For the 1990 to 2000 extrapolation, we had the advantage of an abbreviated 1995 US life table [20], containing single years of age up to age 85, by sex and race. Again, the log-hazard scale seemed most useful, in terms of plots having the smallest variation on this scale. For each race (total, white, non-white, black) and sex, a smoothed fit

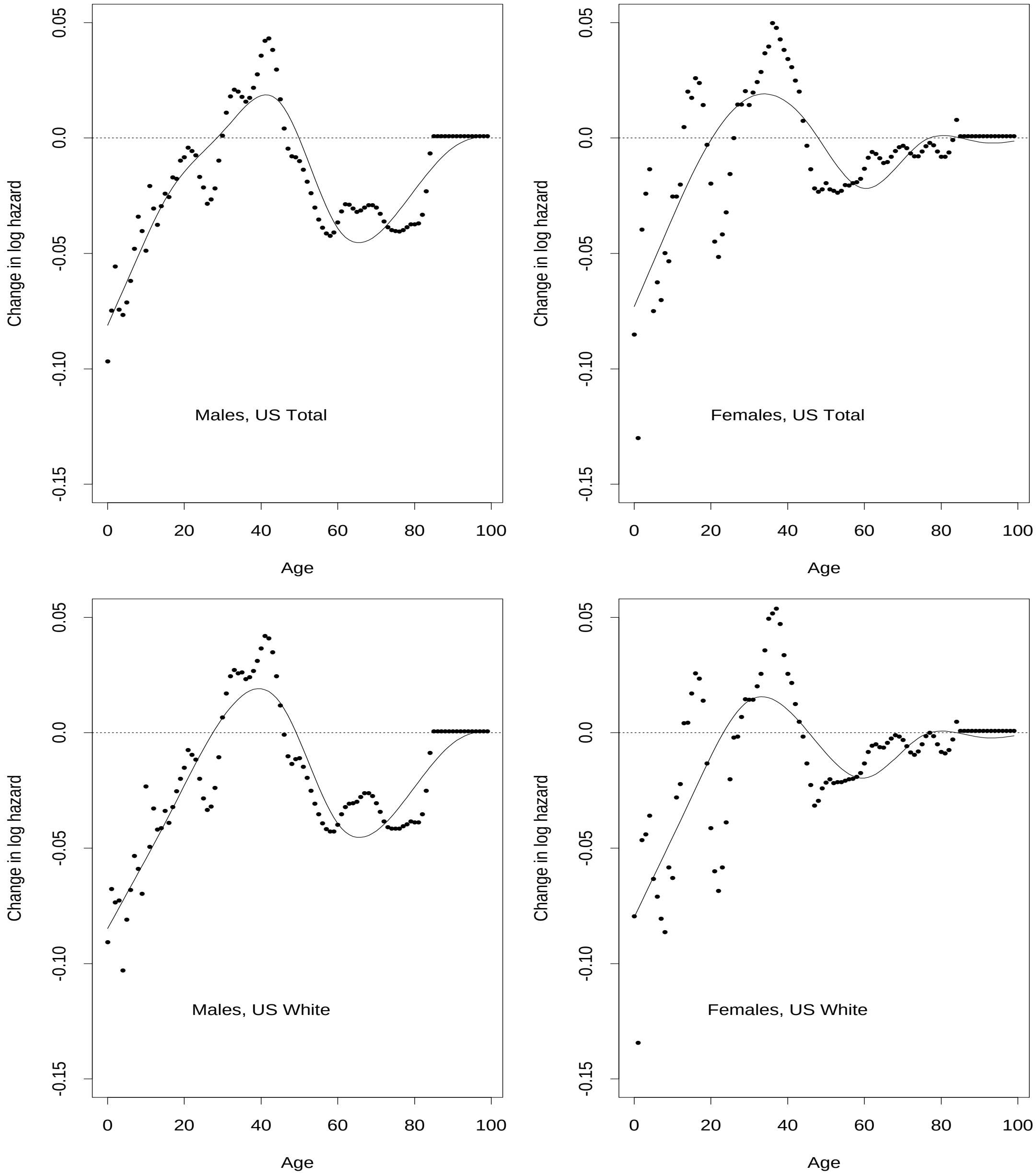

Figure 3: *Actual and smoothed change in hazard, 1990 to 1995*

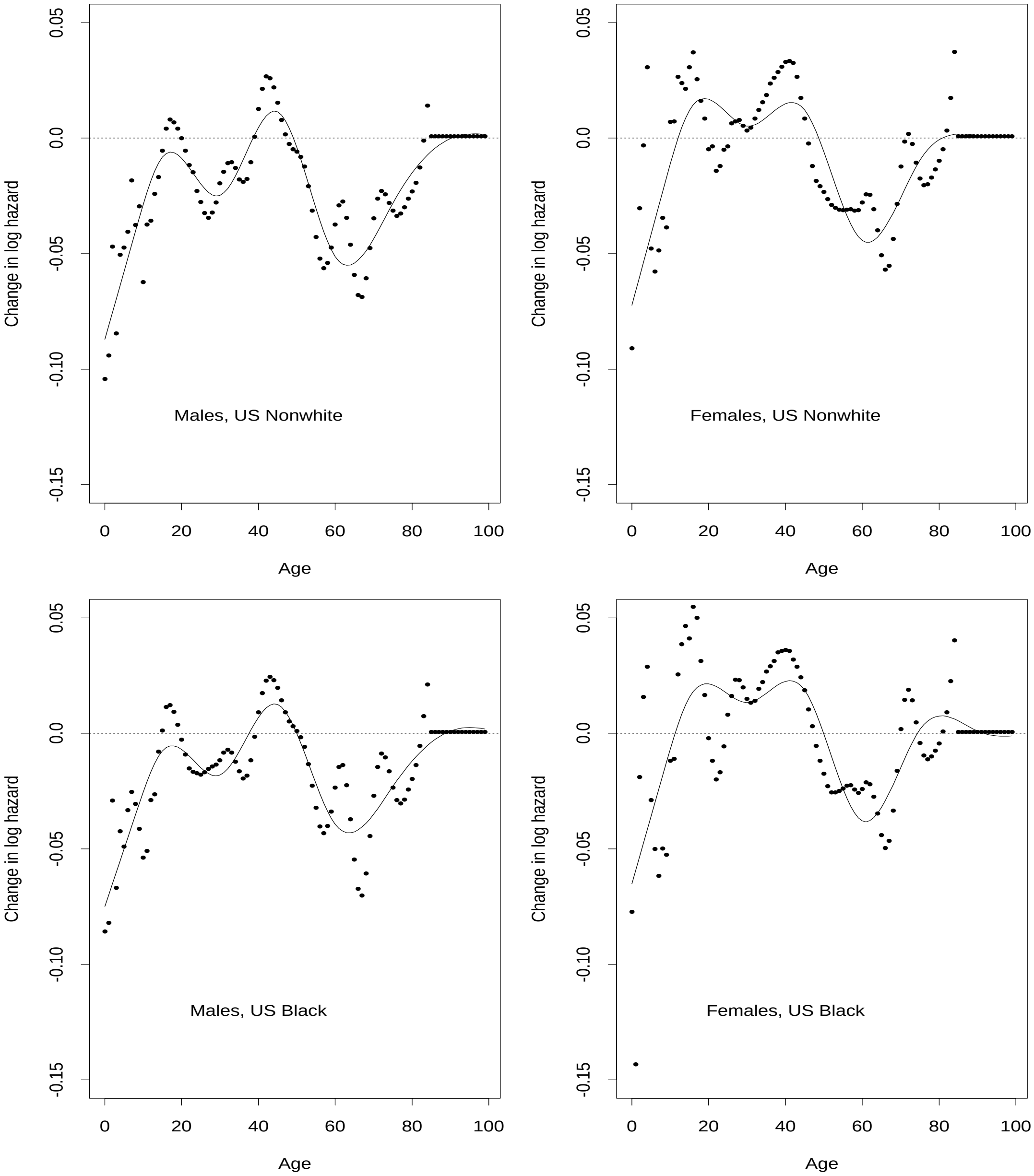

Figure 4: *Actual and smoothed change in hazard, 1990 to 1995*

to the 1990 and 1995 data was obtained as a natural spline with knots at ages 8, 15, 30, 45, 60, 75, 90 and 105, specifically by using the SAS macro `daspline`. The curve is purposely oversmoothed. The smoothing is not as extreme as the prior extrapolation, which assumed that the curve was a constant! The increase in hazard for ages over 100 is largely due to a methodologic change in the way these estimates are computed by the National Center for Health Statistics. The 1995 data was available only through age 85. Because the gains above this age for both men and women, as compared to 1970 rates, were essentially zero, the year 2000 rates for ages 85+ are set equal to the 1990 data. Other than for infants the survival for women has changed very little from 1990 to 1995.

4 Individual Expected Survival

4.1 Population rate tables

In the published life tables, each entry is the probability that a given subject, in a given calendar year, will reach his/her next birthday [17]. The entry for a 20 year old male in 1950, for instance, contains the probability that a subject who turns 20 years of age in 1950 will reach his 21st birthday. The log of this survival probability p_i is related to the cumulative hazard $\Lambda(t)$

$$\log(p_i) = \Lambda(i+1) - \Lambda(i)\,.$$

Assuming that the cumulative hazard is linear over each interval, each subject's cumulative hazard curve is a piecewise linear function with 'elbows' at each birthday. as depicted in figure 5 for a subject born on 11/9/1931.

The table of U.S. hazards has data only for the decades 1960, 1970, etc. Linear interpolation is used for intervening years, e.g. the 1962 value is .8∗(1960 value) + .2∗(1970 value). The rates for the earliest available calendar year are used for all years before this year and the rates for the latest calendar year in the table are used for all years after that year. The rates for the oldest age (109) are used for all subsequent ages.

For integer years of follow up the total survival for a subject can be expressed either using hazards as $\exp(\Lambda(t))$ or as a product of conditional yearly probabilities $\prod p_i$, the two forms give identical answers. For partial years of follow-up the interpolation can be done either on the hazard scale (i.e. as in the figure above) or on the survival scale. The computer functions use the hazard scale because it is easier to deal with partial years.

In detail, the hazard based computation is as follows: we assume that each subject experiences a daily hazard of h_0/day over the first year of life, h_1/day over the second year, ... The cumulative hazard $\Lambda(t)$ is the sum of the daily hazards, and the expected survival at time t is $\exp(-\Lambda(t))$. The major advantage of the cumulative hazard

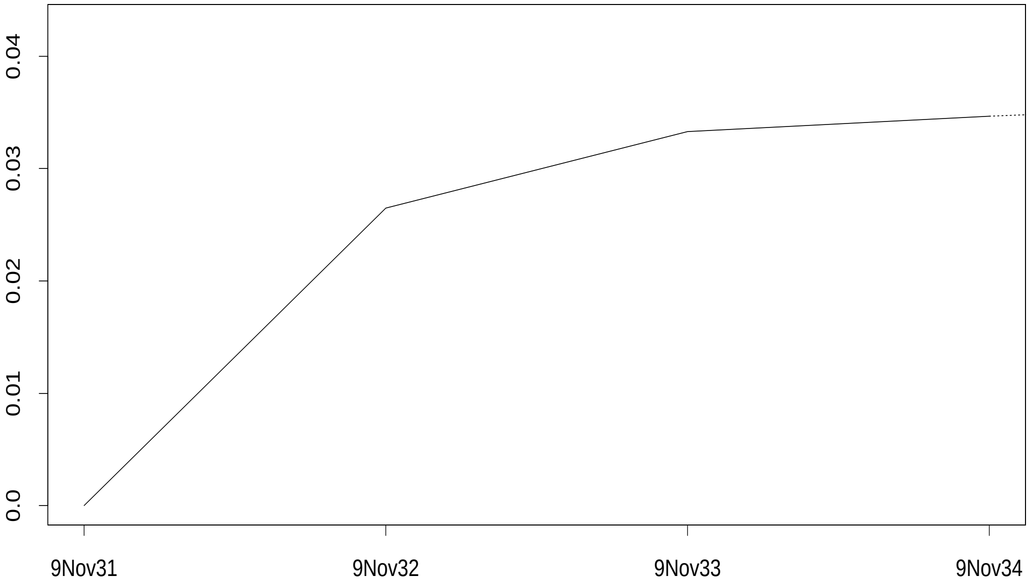

Figure 5: *Cumulative Hazard is piecewise linear over calendar time.*

formulation, as opposed to multiplying the conditional probabilities, is that it is much easier to deal with partial years of follow-up. For example, a woman born on 8/31/42 enters a study on 5/10/63. What is the expected 1 and 2 year survival? The subject is 20 years old on 8/31/62. From the US white female table for 1960, the conditional probability of surviving from the 20th to 21st year is 0.999434 and the corresponding hazard per day is $-\log(.999434)/365.24 = .0000015550$. In 1970, the values are .999355 and 0.0000017724, respectively. Using linear interpolation on the hazard scale, the 1962 hazard rate would be: .8 *(1960 value) + .2*(1970 value) = .0000015985. In like fashion, the hazard from her 21st to 22nd birthday would be .0000016410. Using the hazard formulation, her cumulative hazard for the first year is 10^{-6} times

$$\begin{array}{lclcl} 5/10/63 \text{ to } 8/30/63 & = & 113 \text{ days @ } 1.5985 & = & 180.628 \\ 8/31/63 \text{ to } 5/9/64 & = & 253 \text{ days @ } 1.6410 & = & 415.165 \end{array}$$

So, the 1 year probability of survival is exp (-.0005960) = .9994044. (rounded numbers are printed here, but the computations used exact values).

Using the linear interpolation on the survival scale, as was found in SAS SURVFIT procedure, the survival using the event rates would be computed from the 2 yearly survival rates of exp(-365.24*.0000015985)= .9994163 and .9994008 as

$$\begin{array}{lcccc} 5/10/63 \text{ to } 8/30/63 & = & 1-(113/365)(1-.9994163) & = & .999819 \\ 8/31/63 \text{ to } 5/9/64 & = & 1-(253/365)(1-.9994008) & = & .999585 \end{array}$$

which are multiplied together to obtain an overall survival of .9994041. The numeral difference between the two methods is trivial, but the hazard calculation is more convenient since it is a simple sum.

There are two reasons for using 365.24 instead of 365.25 in our calculations. First, there are 24 leap years per century, not 25. Second, the use of .25 led to some confusing S results when we did detailed testing of the functions, because the S-Plus `round` function uses a nearest even number rule, i.e., `round(1.5) = round(2.5) =2`. In actual data, of course, this niggling detail won't matter a bit.

4.2 User created rate tables

The US and state population tables are somewhat special, in that many other sources for rate data are reported not as a probability of survival p but as $r =$ events per 100,000 subjects per year. The daily hazard table for the computer program could, presumably, be created using either one of these two formulae:

$$-\log(1-10^{-5}r)/365.24$$

or

$$10^{-5}r/365.24\,.$$

For rare events, these two forms will give nearly identical answers. For larger rates, the proper choice depends on whether the rate is computed over a population that is static and therefore depleted by the events in question, or a population that is dynamic and therefore remains approximately the same size over the interval. The first case applies to the standard rate tables, the second may more often apply in epidemiology.

An example rate table is given in section 6.

5 Cohort Expected Survival

The prior section discussed the computation of an expected survival for an individual, here we outline how these are combined to give an overall expected survival for the group. There are several different methods. The various papers in which they are described can be somewhat difficult to compare because they are confounded with different approximation methods for the individual curves, i.e., the subject of the last section.

Let $\lambda_i(t)$ be the expected hazard function for subject i, drawn from a population table, and matched with subject i based on age, sex, and whatever. Then

$$S_i(t) \quad = \quad \exp(-\Lambda_i(t))$$

$$\Lambda_i(t) = \int_0^t \lambda_i(s) ds$$

are the expected cumulative hazard and expected survival curves, respectively, for a hypothetical subject who matches subject i at the start of follow up. For simplicity in some later expressions, also define $h_i(t,s) = \Lambda_i(t+s) - \Lambda_i(t)$, the total hazard accumulated by subject i from time t to time $t+s$.

The expected cumulative hazard and survival for the combined cohort of subjects $i = 1, \ldots, n$ are defined as

$$\begin{aligned}
\Lambda_e(t) &= \int_0^t \frac{\sum_{i=1}^n \lambda_i(s) w_i(s)}{\sum_{i=1}^n w_i(s)} ds \\
S_e(t) &= \exp[-\Lambda_e(t)] \, ,
\end{aligned}$$

where $w_i(t)$ depends on the method. Suggested choices for w are

the *exact* method of Ederer, Axtell and Cutler [5]

$$w_i(t) = S_i(t), \tag{1}$$

the *cohort* method of Hakulinen and Abeywickrama [7]

$$w_i(t) = S_i(t) c_i(t), \tag{2}$$

the *conditional* estimate of Ederer and Heise [6]

$$w_i(t) = Y_i(t).$$

5.1 The Exact Method

This is perhaps the most intuitive way to weight the expected hazards. The term under the integral is the average of the hazards at time s, and the weights are the probability of a subject being alive at that time. It is thus an average over those still expected to be alive. The exact method gives the survival curve of a fictional matched control group, assuming complete follow-up for all of the controls. This is perhaps easier to see if we rewrite the formula as

$$\begin{aligned}
S_e(t) &\equiv \exp(-\Lambda_e(t)) \\
&= \exp\left(\int_0^t \left[\frac{\partial}{\partial u} \log\{(1/n) \sum_{i=1}^n S_i(u)\}\right] du\right) \\
&= (1/n) \sum_{i=1}^n S_i(t) \, . \qquad (3)
\end{aligned}$$

Equation (3) is the usual definition of the exact method. It is interesting to note that in the paragraph just above this definition ([5] page 110), the verbal description of the method suggests an average over those who actually survive to time t, which is the conditional estimate of Ederer and Heise. A third expression, and the form actually used by the program, is easily derived from the above.

$$S_e(t+s) = S_e(t) \, \frac{\sum w_i(t) e^{-h_i(t,s)}}{\sum w_i(t)} \, , \tag{4}$$

where $w_i(t) \equiv S_i(t)$. This gives the total survival as a product of conditional survivals.

One technical problem with the exact method is that it often requires population data that is not yet available. For instance assume that a study is open for enrollment from 1980 to 1990, with follow-up to the analysis date in 1993. If a 11 year expected survival were produced on 1/93, the *complete* expected follow-up data for the last subject enrolled involves the year 2001 US population data.

5.2 The cohort method

Several authors have shown that the Ederer method can be misleading if censoring is not independent of age and sex (or whatever the matching factors are for the referent population). Indeed, independence is often not the case. In a long study it is not uncommon to allow older patients to enroll only after the initial phase. A severe example of this is demonstrated in Verhuel et al. [15], concerning aortic valve replacement over a 20 year period. The proportion of patients over 70 years of age was 1% in the first ten years, and 27% in the second ten years. Assume that analysis of the data took place immediately at the end of the study period. Then the Kaplan-Meier curve for the latter years of follow-up time is guaranteed to be "flatter" than the earlier segment, because it is computed over a much younger population. The Ederer curve will not reflect this bias in the K-M, and give a false impression of utility for the treatment.

In Hakulinen's method [7, 8], each study subject is again paired with a fictional referent from the cohort population, but this referent is now treated as though he/she were followed-up in the same way as the study patient. Each referent is thus exposed to censoring, and in particular has a maximum *potential* follow-up, i.e., they will become censored at the analysis date. In the Hakulinen weight (equation 2), c_i is a censoring indicator which is 1 during the period of potential follow-up and 0 thereafter. If the study subject is censored then the referent would presumably be censored at the same time, but if the study subject dies the censoring time for his/her matched referent will be the time at which the study subject *would have been censored.* For observational studies or clinical trials where censoring is induced by the analysis date this should be straightforward, but determination of the potential follow-up could be a problem if there are large numbers lost to follow-up. (However, as pointed out long ago by Berkson, if a large number of subjects are lost to follow-up then any conclusion is subject to doubt).

In practice, the program can be invoked using the actual follow-up time for those patients who are censored, and the *maximum* potential follow-up for those who have died. By the maximum potential follow-up we mean the difference between enrollment date and the most optimistic last contact date, e.g., if patients are contacted every 3 months on average and the study was closed six months ago this date would be 7.5 months ago. It may be true that the (hypothetical) matched control for a case who died 30 years ago would have little actual chance of such long follow-up, but this is not really important. Almost all of the numerical difference between the exact and cohort estimates results from censoring those patients who were most recently entered on study.

Assume that for some time interval $(t, t+s)$ the weights $w_i(\cdot)$ are constant for all i, i.e., that the potential risk set remains constant over the interval. Then using the same manipulation as in equation (3), equation (4) is found to hold for the cohort estimate as well, with $S_i(t)c_i(t)$ as the weights. This is the estimator used by the program.

This formula differs somewhat from that presented in Hakulinen [8]. He assumes that the data are grouped in time intervals, and thus develops a modification of the usual actuarial formula. The numerical difference, however, should be trivial if the midpoints of these grouped intervals were used in (4).

5.3 Conditional Expected Survival

The conditional estimate is advocated by Verhuel [15], and was also suggested as a computation simplification of the exact method by Ederer and Heise [6]. The weight $Y_i(t)$ is 1 if the subject is alive and at risk at time t, and 0 otherwise. The estimate is clearly related to Hakulinen's cohort method, since $E(Y_i(t)) = S_i(t)c_i(t)$. However, when considered as a product of conditional estimates, it's form is somewhat different than (4); in this case

$$S_e(t+s) = S_e(t)\ \exp\left(-\frac{\sum h_i(t,s)Y_i(t)}{\sum Y_i(t)}\right) . \tag{5}$$

As for the cohort estimate, the derivation requires that $Y_i(\cdot)$ be constant over the interval $(t, t+s)$, i.e., no one dies or is censored in the interior of the interval.

One advantage of the conditional estimate, shared with Hakulinen's method, is that it remains consistent when the censoring pattern differs between age-sex strata. This advantage was not noted by the Ederer and Heise, and the "exact" calculation was adapted as the preferred method [5, 7]. A problem with the conditional estimator is that it has a much larger variance than either the exact or cohort estimate. In fact, the variance of these latter two can usually be assumed to be zero, at least in comparison to the variance of the Kaplan-Meier of the sample. Rate tables are normally based on a very large sample size so the individual rates λ_i are very precise, and the censoring indicators $c_i(t)$ are based on the the study design rather than on patient outcomes. The conditional estimate of $S_e(t)$, however, depends on the observed survival up to t.

5.4 Recommendation

Because it predicts the outcome of a hypothetical group at the completion of their follow-up, the Ederer curve is the most natural to use for study planning activities such as sample size. If the expected survival curve is going to be compared to the observed (K-M) survival curve, either graphically or numerically, then the exact method should not be used unless there is convincing evidence that censoring is unrelated to any of the factors (age, sex, etc.) used to match the study group to the referent population. Such evidence is difficult to come by. It remains the easiest calculation to do by hand, but computer programs would seem to have made this advantage irrelevant.

The conditional estimate is the next easiest to compute, since it requires only the follow-up time and status indicators necessary for the Kaplan-Meier. The actual curve generated by the conditional estimator remains difficult to interpret, however. One wag in our department has suggested calling it the "lab rat" estimator, since the control subject is removed from the calculation ("sacrificed") whenever his/her matching case dies. Andersen and Væth make the interesting suggestion that the difference between the log of the conditional estimate and the log of the Kaplan-Meier can be viewed as an estimate of an additive hazard model

$$\lambda(t) = \lambda_e(t) + \alpha(t) \, ,$$

where λ is the hazard for the study group, λ_e is the expected hazard for the subjects and α the excess hazard created by the disease or condition. Thus the difference between curves may be interpretable even though the conditional estimate $S_e(t)$ itself is not.

We suggest that Hakulinen's cohort estimate is the most appropriate for common use, and particularly for any graphical display alongside of the Kaplan-Meier of the data.

5.5 Approximations

The above equations (4) and (5) are "Kaplan-Meier like" in that they are a product of conditional probabilities and that the time axis is partitioned according to the observed death and/or censoring times. They are unlike a KM calculation, however, in that the ingredients of each conditional estimate are the n distinct individual survival probabilities at that time point rather than just a count of the number at risk. For a large data set this requirement for $O(n)$ temporary variables may be a problem, particularly for the SAS macro. An approximation is to use longer fixed width intervals, and allow subjects to contribute partial information to each interval. For instance, in (5) replace the 0/1 weight $Y_i(t)$ by $\int_t^{t+s} Y_i(u)du/s$, which is the proportion of time that subject i was at risk during the interval $(t, t+s)$. A similar proportionality correction can be made to the weights in equation (4) for the cohort estimate: $c_i(t)$ is replaced by the proportion of time that subject i was uncensored during the interval $(t, t+s)$.

If those with fractional weights form a minority of those at risk during the interval the approximation should be reliable. (More formally, if the sum of their weights is a minority of the total sum of weights). By Jensen's inequality, the approximation will always be biased upwards. However, the bias is usually very small. For the Stanford heart transplant data used in the examples below an exact 5 year estimate using the cohort method is 0.94728, a computation using half year intervals yields 0.94841. Even with these very wide intervals the difference is only in the third decimal place.

The Ederer estimate is unchanged under repartitioning of the time axis.

5.6 Total expected deaths

All of the above discussion has been geared towards a plot of $S_e(t) = \exp(-\Lambda_e(t))$, which attempts to capture the proportion of patients who will have died by t. When comparing observed to expected survival for testing purposes, an appropriate test is the one-sample logrank test $(O-E)^2/E$ [10], where O is the observed number of deaths and

$$\begin{aligned} E &= \sum_{i=1}^n e_i \\ &= \sum_{i=1}^n \int_0^\infty \lambda_i(s) Y_i(s)\, ds \qquad (6) \end{aligned}$$

is the expected number of deaths, given the observation time of each subject. This follows Mantel's concept of 'exposure to death' [11], and is the expected number of deaths during this exposure. Notice how this differs from the expected number of deaths in the matched cohort at time t: $nS_e(t)$. In particular, E can be greater than n. The SAS `ltp` macro and the S `survexp` function (with the `cohort=F` option) both return the individual expected survivals $\exp(-e_i)$.

Equation (6) is referred to as the person-years estimate of the expected number of deaths. The logrank test is usually more powerful than one based on comparing the observed number of deaths by time t to $nS_e(t)$; the former is a comparison of the entire observed curve to the expected, and the latter is a test for difference at one point in time.

Tests at a particular time point, though less powerful, will be appropriate if some fixed time is of particular interest, such as 5 year survival. In this case the test should be based on the cohort estimate. The H_0 of the test is "is observed survival at t the same as a control-group's survival would have been". A pointwise test based on the conditional estimate has two problems. The first is that an appropriate variance is more difficult to construct. The second, and more damning one, is that it is unclear exactly what alternative is being tested against.

Berry [3] shows how the individual expected hazards e_i may be used to adjust regression models. The one-sample logrank test is seen to be equivalent to the test for `intercept=0` in a Poisson model with $\log(e_i)$ as an offset term, replacing the usual offset of $\log(t_i)$. This may be extended to more complicated regression models, e.g., to compare the excess death rates among multiple groups. An offset of $\log(e_i)$ may also be used in a Cox model, to correct for differential background mortality.

6 S Implementation

The rate tables are used by the S-Plus `survexp` and `pyears` functions to obtain expected survival and person-years computations, respectively. As a first example, we will calculate the expected survival for the Stanford heart transplant data set, as found in the JASA article of Crowley and Hu [4]. This data set contains birth, entry, and last follow-up dates, treatment, and prior surgery as covariates. Sex will be assumed to be male, and we will use the US total population as the comparison data set. The last potential follow-up date for any subject was April 1 1974. A copy of the data set can be found on Statlib. The following code will calculate the Ederer or "exact" estimate, with separate curves for the two treatment arms.

```
# exact estimate
attach(jasa)
rx  <- !is.na(tx.date)
age <- (entry.dt - birth.dt)   # age in days
exp1 <- survexp( ~ rx + ratetable(age=age, year=entry.dt, sex=1),
                 data=jasa, ratetable=survexp.us, times=(0:4)*182.5)
```

The `ratetable` function is used to match the data set's variable names to the `age`, `sex` and `year` dimensions of the US table. The arguments to `ratetable` can be in any order. If the input data contains the same variable names (with the correct coding!) as the rate table, then the `ratetable` function is not needed. That is, an alternative to the above code is:

```
mydata <- data.frame(jasa, age=jasa$entry.dt - jasa$birth.dt, sex=1,
                      year=jasa$entry.dt)
exp1 <- survexp( ~ rx , data=mydata,
                        ratetable=survexp.us, times=(0:4)*182.5)
```

The times argument specifies that an output estimate should be computed at half year intervals for 2 years. The resultant curves can be listed or drawn using print and plot functions.

The cohort estimate uses potential follow-up on the left hand side, along with the conditional argument. The potential follow-up time for a censored subject is the observed follow-up time, but for someone who dies it is the amount of time they might have been followed had the death not occurred.

```
# cohort estimate
ptime <- mdy.date(4,1,74) - entry.dt
ptime <- ifelse(fustat==1, ptime, futime)
exp3 <- survexp( ptime ~ rx, data=mydata, ratetable=survexp.us,
                ratetable=survexp.us, times=(0:4)*182.5, conditional=F)
```

If the times argument is omitted, an estimate is returned for each unique follow-up time.

To compute the conditional estimate, follow-up time is included on the left hand side of the formula.

```
# conditional estimate
futime <- fu.date - entry.dt
exp2 <- survexp( futime ~ rx, data=mydata, conditional=T,
                 ratetable=survexp.us, times=(0:4)*182.5)
```

By default, the survexp function returns a survival curve for the entire cohort of subjects. To use expected survival as a covariate in a model a single number per subject is desired, i.e., the subjects' expected hazard on their last follow up date. For instance, the following computes the one sample logrank test (the test for intercept=0 in fit1) and a test for treatment difference after controlling for baseline mortality due to age (the test for rx=0 in fit2). Note the argument cohort=F. The vector haz will contain the individual values e_i of equation (6).

```
# individual expected survival
haz <- -log(survexp(futime ~ 1, data=mydata,
                      ratetable=survexp.us, cohort=F))
fit1 <- glm(fustat ~ offset(log(haz)), data=jasa, family=poisson)
fit2 <- glm(fustat ~ rx + offset(log(haz)), data=jasa, family=poisson)
```

By default the internal computations used in survexp partition the time line at every censoring or death point, thus equations (4) and (5) hold exactly. For very large data sets the npoints option may be used to replace this with the approximation discussed in section 4.4.

User created rate tables may be used in place of the provided populations. Tables 1 and 2 show yearly death rates per 100,000 subjects based on their smoking status [21]. A stored raw data set contains this data, with the "Never smoked" data replicated

			Former smokers (1-20 cig/day)					
			Duration of abstinence (yr)					
	Never	Current						
Age	Smoked	Smokers	< 1	1-2	3-5	6-10	11-15	≥ 16
45-49	186.0	439.2	234.4	365.8	159.6	216.9	167.4	159.5
50-54	255.6	702.7	544.7	431.0	454.8	349.7	214.0	250.4
55-59	448.9	1,132.4	945.2	728.8	729.4	590.2	447.3	436.6
60-64	733.7	1,981.1	1,177.7	1,589.2	1,316.5	1,266.9	875.6	703.0
65-60	1,119.4	3,003.0	2,244.9	3,380.3	2,374.9	1,820.2	1,669.1	1,159.2
70-74	2,070.5	4,697.5	4,255.3	5,083.0	4,485.0	3,888.7	3,184.3	2,194.9
75-79	3,675.3	7,340.6	5,882.4	6,597.2	7,707.5	4,945.1	5,618.0	4,128.9

			Former smokers (≥ 21 cig/day)					
			Duration of abstinence (yr)					
	Never	Current						
Age	Smoked	Smokers	< 1	1-2	3-5	6-10	11-15	≥ 16
45-49		610.0	497.5	251.7	417.5	122.6	198.3	193.4
50-54		915.6	482.8	500.7	488.9	402.9	393.9	354.3
55-59		1,391.0	1,757.1	953.5	1,025.8	744.0	668.5	537.8
60-64		2,393.4	1,578.4	1,847.2	1,790.1	1,220.7	1,100.0	993.3
65-69		3,497.9	2,301.8	3,776.6	2,081.0	2,766.4	2,268.1	1,230.7
70-74		5,861.3	3,174.6	2,974.0	3,712.9	3,988.8	3,268.6	2,468.9
75-79		6,250.0	4,000.0	4,424.8	7,329.8	6,383.0	7,666.1	5,048.1

Table 1: *Deaths per 100,000/year, males*

where the lower table shows blanks, followed by the data for females. A rate table is created using the following S code.

```
temp <- matrix(scan("data.smoke"), ncol=8, byrow=T)/100000
smoke.rate <- c(rep(temp[,1],6), rep(temp[,2],6), temp[,3:8])
attributes(smoke.rate) <- list(
    dim=c(7,2,2,6,3),
    dimnames=list(c("45-49","50-54","55-59","60-64","65-69","70-74","75-79"),
                  c("1-20", "21+"),
                  c("Male","Female"),
                  c("<1", "1-2", "3-5", "6-10", "11-15", ">=16"),
                  c("Never", "Current", "Former")),
    dimid=c("age", "amount", "sex", "duration", "status"),
    factor=c(0,1,1,0,1),
    cutpoints=list(c(45,50,55,60,65,70,75),NULL, NULL,
                                 c(0,1,3,6,11,16),NULL),
    class='ratetable'
    )
is.ratetable(smoke.rate)
```

The smoking data cross-classifies subjects by 5 characteristics: age group, sex, status (never, current or former smoker), the number of cigarettes consumed per day, and, for the prior smokers, the duration of abstinence. In our S implementation, a ratetable is an array with added attributes, and thus must be rectangular. In order to cast the above data into a single array, the rates for never and current smokers needed to be replicated across all 6 levels of the duration, we do this by first creating the `smoke.rate` vector. The array of rates is then saddled with a list of descriptive attributes. The dim and dimnames are as they would be for an array, and give its shape and printing labels, respectively. Dimid is the list of keywords that will be recognized by the `ratetable` function, when this table is later used within the `survexp` or `pyears` function. For the US total table, for instance, the keywords are “age”, “sex”, and “year”. These keywords must be in the same order as the array dimensions (as found in the dimid attribute, not in the user invocation). The factor attribute identifies each dimension as fixed or mobile in time. For a subject with 15 years of follow-up, for instance, the sex category remains fixed over this 15 years, but the age and duration of abstinence continue to change; more than 1 of the age groups will be referenced to calculate his/her total hazard. For each dimension that is not a factor, the starting value for each of the rows of the array must be specified so that the routine can change rows at the appropriate time, this is specified by the cutpoints. The cutpoints are null for a factor dimension. Because these attributes must be self-consistent, it is wise to carefully check them for any user created rate table. The `is.ratetable` function does this automatically.

As a contrived example, we can apply this table to the Stanford data, assuming that all of the subjects were current heavy smokers (after all, they have heart disease).

			Former smokers (1-20 cig/day)					
			Duration of abstinence (yr)					
	Never	Current						
Age	Smoked	Smokers	< 1	1-2	3-5	6-10	11-15	≥ 16
45-49	125.7	225.6		433.9	212.0	107.2	135.9	91.0
50-54	177.3	353.8	116.8	92.1	289.5	200.9	121.3	172.1
55-59	244.8	542.8	287.4	259.5	375.9	165.8	202.2	247.2
60-64	397.7	858.0	1,016.3	365.0	650.9	470.8	570.6	319.7
65-60	692.1	1,496.2	1,108.0	1,348.5	1,263.2	864.8	586.6	618.0
70-74	1,160.0	2,084.8	645.2	1,483.1	1,250.0	1,126.3	1,070.5	1,272.1
75-79	2,070.8	3,319.5		2,580.6	2,590.7	3,960.4	1,666.7	1,861.5

			Former smokers (≥ 21 cig/day)					
			Duration of abstinence (yr)					
	Never	Current						
Age	Smoked	Smokers	< 1	1-2	3-5	6-10	11-15	≥ 16
45-49	125.7	277.9	266.7	102.7	178.6	224.7	142.1	138.8
50-54	177.3	517.9	138.7	466.8	270.1	190.2	116.8	83.0
55-59	244.8	823.5	473.6	602.0	361.0	454.5	412.2	182.1
60-64	397.7	1,302.9	1,114.8	862.1	699.6	541.7	373.1	356.4
65-69	692.1	1,934.9	2,319.6	1,250.0	1,688.0	828.7	797.9	581.5
70-74	1,160.0	2,827.0	4,635.8	2,517.2	1,687.3	2,848.7	1,621.2	1,363.4
75-79	2,070.8	4,273.1	2,409.6	5,769.2	3,125.0	2,978.7	2,803.7	2,195.4

Table 2: *Deaths per 100,000/year, females*

```
# user supplied rate table
p2 <- ptime/365.24
exp4 <- survexp(p2 ~ ratetable(age=(age/365.24), status="Current",
                                   amount=2, duration=1, sex='Male'),
            data=jasa, ratetable=smoke.rate, conditional=F, scale=1)
```

This example does illustrate some points. For any factor variable, the `ratetable` function allows use of either a character name or the actual column number. Since I have chosen the current smoker category, duration is unimportant, and any value could have been specified. The most important point is to note that `age` has been rescaled. This table contains rates per year, whereas the US tables contained rates per day. It is crucial that all of the time variables (age, duration, etc) be scaled to the same units, or the results may not be even remotely correct. The US rate tables were created using days as the basic unit since year of entry will normally be a julian date; for the smoking data years seemed more natural.

An optional portion of a rate table, not illustrated in the example above, is a `summary` attribute. This is a user written function which will be passed a matrix and can return a character string. The matrix will have one column per dimension of the ratetable, in the order of the `dimid` attribute, and will have already been processed for illegal values. To see an example of a summary function, type `attr(survexp.us, 'summary')` at the S prompt. In this summary function the returned character string lists the range of ages and calendar years in the input, along with the number of males and females. This string is included in the output of `survexp`, and will be listed as part of the printed output. This printout is the only good way of catching errors in the time units; for instance, if the string contained "age ranges from .13 to .26 years", it is a reasonable guess that age was given in years when it should have been stated in days.

The data could have been organized in other ways, for instance as a 2 by 7 by 15 array based on sex, age, and a 15 level grouping variable with levels "Never smoked", "Current smoker of 1-20 cig/day", "Current smoker of > 20 cig/day", "Former smoker of 1-20 but ceased for < 1 year", ...

As an aside, many entries in the smoke.rate table are based on small samples. In particular, the data for females who are former smokers contains 2 empty cells. Before serious use these data should be smoothed. As a trivial example:

```
newrate <- smoke.rate
temp <- newrate[ ,1,2, ,3]
fit <- gam(temp ~ s(row(temp)) + s(col(temp)))
newrate[,1,2,,3] <- predict(fit)
```

A realistic effort would begin and end with graphical assessment, and likely make use of the individual sample sizes as well.

References

[1] Andersen, P. and Væth, M. (1989). Simple parametric and nonparametric models for excess and relative mortality. *Biometrics* **45**, 523-35.

[2] Bergstralh, E. and Offord, K.(1988). Conditional probabilities used in calculating cohort expected survival. *Technical Report #37*, Section of Medical Research Statistics, Mayo Clinic.

[3] Berry, G. (1983). The analysis of mortality by the subject-years method. *Biometrics* **39**, 173-84.

[4] Crowley, J. and Hu, M. (1977), Covariance analysis of heart transplant data. *J. Am. Stat. Assoc.* **72**, 27-36.

[5] Ederer, F., Axtell, L.M. and Cutler, S.J. (1961). The relative survival rate: a statistical methodology. *National Cancer Inst Monographs* **6**, 101-21.

[6] Ederer, F. and Heise, H. (1977). Instructions to IBM 650 programmers in processing survival computations, *Methodological Note No. 10, End Results Evaluation Section, National Cancer Institute.*

[7] Hakulinen, T. (1982). Cancer survival corrected for heterogeneity in patient withdrawal. *Biometrics* **38**, 933.

[8] Hakulinen, T. and Abeywickrama, K.H. (1985). A computer program package for relative survival analysis. *Computer Programs in Biomedicine* **19**, 197-207.

[9] Hakulinen, T. (1977). On long term relative survival rates. *J. Chronic Diseases* **30**, 431-43.

[10] Harrington, D.P. and Fleming, T.R. (1982). A class of rank test procedures for censored survival data. *Biometrika*, **69**, 553-66.

[11] Mantel, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemotherapy Reports* **50**, 163-6.

[12] Offord, K.; Augustine, G.; Fleming, T.; and Scott, W.(1986) *The SURVFIT Procedure.* SUGI Supplemental Library User's Guide, Version 5, SAS Institute Inc., Cary, NC.

[13] Therneau, T., Sicks, J., Bergstralh, E. and Offord, J. (1994). Expeceted Survival Based on Hazard Rates, *Technical Report No. 54*, Department of Health Science Research, Mayo Clinic.

[14] Therneau, T. and Scheib, C. (1994). Extrapolation of the U.S. Life Tables, *Technical Report No. 55*, Department of Health Science Research, Mayo Clinic.

[15] Verhuel, H.A., Dekker, E., Bossuyt, P., Moulijn, A.C. and Dunning, A.J. (1993). Background mortality in clinical survival studies. *Lancet* **341**, 872-5.

[16] National Center for Health Statistics. Vital statistics of the United States, 1991, vol II, mortality, part A. Washington, Public Health Service, 1996.

[17] National Center for Health Statistics: *Life tables for the geographic divisions of the United States: 1959-61.* Vol 1, number 3. Public Health Service, Washington. U.S. Government Printing Office, May 1965.

[18] National Center for Health Statistics: *Life tables for the geographic divisions of the United States: 1989-91.* Vol 1, number 1. Hyattsville, Maryland, 997.

[19] Vital Statistics of the United States, 1960. Volume II, Mortality, Part A, Section 3, table 3-B. US Department of Health, Education, and Welfare, Washington, 1963.

[20] Vital Statistics of the United States, 1995, Life Tables. Preprint of Volume II, Mortality, Part A, Section 6. National Center for Health Statistics, Hyattsville, May 1998.

[21] *The Health Benefits of Smoking Cessation* (1990). US Department of Health and Human Services. Public Health Service, Centers for Disease Control, Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health. DHHS Publication No (CDC)90-8416.