EXPECTED SURVIVAL BASED ON HAZARD RATES

Terry Therneau
JoRean Sicks
Erik Bergstralh
Jan Offord

Technical Report #52
March 1994

Copyright 1994 Mayo Foundation

Expected Survival Based on Hazard Rates

Terry Therneau
JoRean Sicks
Erik Bergstralh
Jan Offord

1 Introduction

This work began in an effort to implement expected survival routines in the S package, similar to the functionality contained in the SAS procedures `survfit` and `survdif`. The tables of survival probabilities used by those two programs form the basis for the calculations. These tables had been compiled over several years by members of the Department of Health Sciences Research, and are documented in [2] and [12]. Initial exploration was focused on plots of this data, in order to understand its structure, and some data anomalies were discovered in this process. The lion's share of the data, as currently used, is discussed in section 2. Further detail on the West North Central data set is in appendix 2. Section 3 shows how these rate tables can be used to construct the survival curve for an individual subject.

Exploration of some recent papers on expected survival for a cohort of subjects has cast doubt on the computational *method* chosen for the original SAS procedures. Section 4 contains an overview and comparison of the competing methods. Sections 5 and 6 discuss new S 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

2.1 Corrections

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 year lag, e.g., we expect to have the 1990 data available by 1995. Each table is based on the average of 3 years, e.g., 1989-1991. The table entry $q_{1960,24,F}$ would contain the probability that a female who became 24 years old sometime in 1960 will die on or before her 25th birthday.

Because of some rounding errors that were discovered, as well as the need to add Florida and Arizona tables, all of the tables were re-entered in January 1994. This will cause some differences between the answers given by the new routines and those obtained with the older SAS `survfit` procedure. As well, some concerns were noted with the West North Central data set, which is derived from a number of sources and would have been quite difficult to recreate "from scratch". These issues and the corrections are detailed in appendix 2.

2.2 Sources

The following are the sources for the US and state tables.

2.2.1 United States

1950 Life Tables for 1949-51, Public Health Service Publication No. ?, Volume ?, No. ? (we have only a photocopy of the relevant pages).

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

2.2.2 West North Central

See table 1 in [2].

2.2.3 Minnesota

1950 Unknown (the files contain a photocopy, without reference).

1960 Minnesota State Lifetables 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

2.2.4 Florida

1970 U.S. Decennial Lifetables 1969-71, DHEW Publication No. HRA 75-1151 , Volume 2, No. 10

1980 U.S. Decennial Lifetables 1979-81, DHEW Publication No. PHS 86-1151-10 , Volume 2, No. 10

2.2.5 Arizona

1970 U.S. Decennial Lifetables 1969-71, DHEW Publication No. HRA 75-1151 , Volume 2, No. 3

1980 U.S. Decennial Lifetables 1979-81, DHEW Publication No. PHS 76-1151-3 , Volume 2, No. 3

2.3 Computer Tables

The following rate tables have been created. All of the tables are by age (0-109), sex and calendar year. Only the decade calendar years are included; intervening year's data is created within the functions by linear interpolation. The S version of the WNC table has separate rates for 0 – 6 months and 6 months – 1 year and yearly ages beyond that. All other tables are by year of age.

2.3.1 S

The tables are part of the survival package, and are automatically attached when Splus is invoked. The master copy for the tables is on the RCF machine *feynman*, along with that for the remainder of the survival functions. Details on the internal format are given in the section of S examples.

`survexp.us` United States total for calendar years 1960 - 1980.

`survexp.uswhite` US white for calendar years 1950-1980.

`survexp.usr` US survival for calendar years 1960-1980, by race. The race groupings are white, non-white, and black. Data for blacks was not available in the 1960 and 1970 data, and the non-white values were used.

`survexp.az` Total survival for Arizona, calendar years 1970-1980.

`survexp.azr` Arizona survival by race, white and non-white, for 1970-80. The non-white values for 1970 were unavailable, and the 1980 non-white was used.

`survexp.fl` Total survival for Florida, calendar years 1970-1980.

`survexp.flr` Florida survival by race, white, non-white and black, for 1970-80. The black values for 1970 were unavailable, and the 1970 non-white was used.

`survexp.mn` Total survival for Minnesota, calendar years 1970-1980.

`survexp.minnwhite` Minnesota white survival for calendar years 1950-80.

`survexp.wnc` West North Central white survival for 1920-1980.

2.3.2 SAS

The recently entered population data is stored on the panvalet archives under member `D1549201` with the following format:

Col 1-2	=	population (MN,US,FL,AZ,WN)
3	=	race (T=total,B=black,W=white, N=non-white)
4	=	sex (M,F)
5-8	=	year (1960,1970,1980)
9-11	=	age (0,109) (whole years only)
12-17	=	proportion dying during year (q) entered without a decimal point

The population data used by SAS are stored in five SAS datasets.

1. `lt_us.ssd01` contains the US population data
2. `lt_mn.ssd01` contains the Minnesota population data
3. `lt_wnc.ssd01` contains the West North Central population data
4. `lt_fl.ssd01` contains the Florida population data
5. `lt_az.ssd01` contains the Arizona population data

Each dataset contains one observation per hazard value, and has the following variables:

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

`year` = decade specification (1910-1980)

`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

These datasets are stored on the following libraries:

1. on unix, under `/usr/local/sasmac`
2. on IBM, on `hsrp.jlk.s50400.pops`

Within the SAS macros using these populations, the following selections are possible for the parameter POP. These specifications use all or portions of the above populations selected in the documented ways:

`us_t` US Total for 1960-1980

`us_w` US White only for 1950-1980

`us_wnw` US White vs Non-white for 1960-1980

`us_bw` US White vs Black for 1980

`wnc_w` West North Central White only for 1910-1980 (note — 1970 and 1980 use MN White rates)

`mn_t` Minnesota Total for 1970-1980

`mn_w` Minnesota White only for 1950-1980

`fl_t` Florida Total for 1970-1980

`fl_w` Florida White only for 1970-1980

`fl_wnw` Florida White vs Non-white for 1970-1980

`fl_bw` Florida White vs Black for 1980

`az_t` Arizona Total for 1970-1980

`az_w` Arizona White only for 1970-1980

`az_wnw` Arizona White vs Non-white for 1980

2.4 Extrapolation

There is a time lag of 4-5 years between each census and the publication of the corresponding rate tables; we expect that the 1990 tables will become available some time in 1994 or 1995. Until then, the last available year's data is used for all follow-up after 1980. A look at figure 7, however, shows that the death rates for most ages have shown a steady decline over the years. The present method, which amounts to extrapolating these curves to the right by horizontal lines, does not appear to be a very wise extrapolation rule.

There are plans to improve this, using a 1994 summer student as the extra labor. When completed year 1990 and 2000 "data" will be added to the rate tables to produce the appropriate extrapolation, without needing to change the S or SAS functions. This will be documented in a later technical report.

3 Individual Expected Survival

3.1 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 [14]. 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, somewhat as depicted in figure 1 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) +

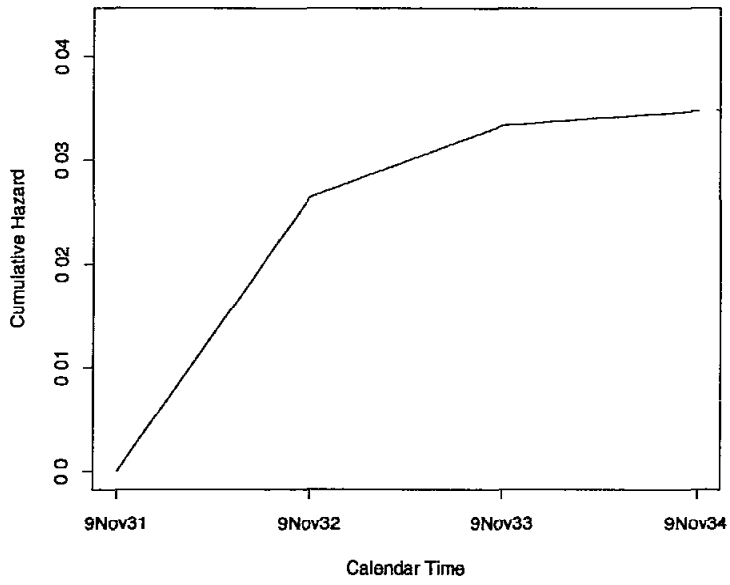

Figure 1: *Cumulative hazard vs calendar time.*

.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 former is used by the new routines described in this paper, the latter was used by `survfit`.

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 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

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

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 found in `survfit`, the survival using the event rates would be computed from the 2 yearly survival rates of exp(-365.24*.0000015985)= .9994163 and .9994008 as

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

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.

A more substantial difference comes from alignment. In the example above, `survfit` would actually not do the calculation that was outlined, but rather it would report the 1962 value of .9994163 as the one year probability of survival. That is, `survfit` acts as though the subject were 20.0 years old on the entry date of 5/10/63. (For the usual program request of 1, 2, etc year expected survival values this considerably simplifies the calculations.) Essentially, we treat patients as being slightly older than `survfit` does. The amount of difference that this makes depends on the patient's age, follow-up time, and the time between enrollment and the last birthday.

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 `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.

3.2 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.

4 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

$$\begin{aligned} S_i(t) &= \exp(-\Lambda_i(t)) \\ \Lambda_i(t) &= \int_0^t \lambda_i(s)ds \end{aligned}$$

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]

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

4.1 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

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

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.

The procedure used in the past by our department, `survfit`, is based on the exact method.

4.2 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. [13], 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 referrent is now treated as though he/she were followed-up in the same way as the study patient. Each referrent 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 referrent would presumably be censored at the same time, but if the study subject dies the censoring time for his/her matched referrent 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 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).

4.3 Conditional Expected Survival

The conditional estimate is advocated by Verhuel [13], 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.

The main argument for use of the conditional estimate is that we often want to make conditional statements about the survival. For instance, in studies of a surgical intervention such as hip replacement, the observed and expected survival curves often will initially diverge due to surgical mortality, and then appear to become parallel. It is tempting to say that survival beyond hospital discharge is "equivalent to expected". This is a conditional probability statement, and it should not be made unless a conditional estimate was used.

A hypothetical example may make this clearer. For simplicity assume no censoring. Suppose we have studies of two diseases, and that their patients' age distributions at entry are identical. Disease A kills 10% of the subjects in the first month, independent of age or sex, and thereafter has no effect. Disease B also kills 10% of its subjects in

the first month, but predominately affects the old. After the first month it exerts a continuing though much smaller force of mortality, still biased toward the older ages. With proper choice of the age effect, studies A and B will have almost identical survival curves; as the patients in B are always younger, on average, than those in A. Two different questions can be asked under the guise of "expected survival":

- What is the overall effect of the disease? In this sense both A and B have the same effect, in that the 5 year survival probability for a diseased group is $x\%$ below that of a matched population cohort. The cohort estimate would be preferred because of its lower variance. It estimates the curve we "would have gotten" if the study had included a control group. Using the cohort method, the expected survival curves for study A and B are identical, which is logical since the hypothetical control groups for the two studies would be identical.

- What is the ongoing effect of the disease? Detection of the differential effects of A and B after the first month requires the conditional estimator. The expected curves computed in this way are not the same; that for disease A will become parallel to the Kaplan-Meier of the group, and that for B would show continued divergence.

Other suggestions for exploring conditional effects can be found in the literature under the heading of relative survival. Hakulinen [9] for instance, suggests dividing the patients into disjoint age groups and computing the ratio of observed/expected survival separately within each strata. However, this estimate can have an unacceptable variance due to small numbers in the subgroups.

4.4 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.

4.5 Recommendation

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. If there is a question about delayed effects the conditional estimator can be used to create a plot of $\alpha(t)$ for inspection. In the example given above, the plots for disease A and B would have a marked change in slope after the first month (the plot for A would, presumably, actually become horizontal). A new Kaplan-Meier and cohort expected curve then could be plotted using only those patients who survived at least one month.

4.6 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

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

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. (As an aside, in his background discussion he neatly summarizes the major issues found in both Hartz et al. and their respondents). 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.

5 S Implementation

The program is implemented as a single S function `survexp()`, along with the rate tables described in section 2. Each of the rate tables is a multi-way array with age, sex, calendar year, and optionally, race as a dimension. As an example, we will calculate 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.

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. 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 + ratetable(age=age, year=entry.dt, sex=1),
                 data=jasa, 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 + ratetable(age=age, year=entry.dt, sex=1),
                 data=jasa, 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 ~ ratetable(age=age, year=entry.dt, sex=1),
        data=jasa, 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. Table 1 shows yearly death rates per 100,000 subjects based on their smoking status [15]. A stored raw data set contains this data, with the "Never smoked" data replicated where the lower table shows blanks, followed by the data for females (female data is not shown in the table for space reasons). 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)
```

Males		Smokers (1-20 cig/day)						
	Never	Current	Former Smokers: Duration of abstinence (yr)					
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

Males		Smokers (≥ 21 cig/day)						
	Never	Current	Former Smokers: Duration of abstinence (yr)					
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 per year based on smoking status*

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. 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 in 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. 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).

```
# 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", ... The SAS example in the next section uses this arrangement.

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 zeros. 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.

6 SAS Implementation

These techniques are implemented in SAS as two macros `%survexp` and `%ltp`, with a third macro `%gethaz` used in the background to read in the populations and compute the linear interpolation. These macros replace the expected portions of the old SAS procedures `survfit` and `survdif`. The `%ltp` macro adds an additional variable `_ltp` to each subject in the input dataset. This variable contains the life table probability that a matched subject would survive to their follow-up time, given age, sex, etc. The `%survexp` macro computes the expected survival curve for the input dataset, using any of the 3 methods presented earlier. Options for computing the one-sample logrank test and/or a graph of the observed vs expected curves are available in `%survexp`. Both macros allow the various populations options described in section 2.3.2, or using a user defined rate table.

Using the same JASA example, the the code to add the _ltp variable to each observation would be:

```
%ltp(pop=US_T, birthdt=birth_dt, firstdt=entry_dt, time=fu_time, data=jasa,
            sex=sex);
```

If you are using a population that is subset by race, the parameter `race` should also be specified. The parameter `firstdt` refers to the entry date or beginning time, `birthdt` references the birth date, and `time` is the observed follow-up time to death or censoring. The dates must be SAS date variables, and the follow-up time must be in days.

To calculate the expected curve values using the exact method (method=1), but not printing or plotting, the statement would be:

```
* Exact method;
%survexp(data=jasa, pop=US_T, birthdt=birth_dt, firstdt=entry_dt,
                   sex=sex, method=1, points=0 to 3650 by 182.5);
```

If you want to print a summary table and one-sample logrank test statistic for this data, you must add the additional parameters `printop`, `plotop`, `time` for the follow-up time, `event` for the event variable, and `cen_vl` for the censoring value. This is similar to `%surv`, and `%surv` is actually called to calculate the observed Kaplan-Meier estimates. When the logrank test is computed `%ltp` is called to compute the expected deaths.

To produce the cohort estimate, the code would look like this:

```
* Cohort method;
if fu_stat=1 then pt_date=mdy(4,1,74);  /* death=fu_stat=1 */
              else pt_date=fu_date;

%survexp(data=jasa, pop=US_T, birthdt=birth_dt, firstdt=entry_dt,
      sex=sex, method=2, lastdt=pt_date);
```

In this example we use the potential follow-up date of 4/1/74 for all the subjects who died, and the actual censoring date for those censored. The parameter `lastdt` containing censoring or potential follow-up date is required for the cohort method. Again, for a summary table and plot the above mentioned additional variables must be added. No `points` parameter was specified, therefore estimates will be calculated for 100 points, (0 to 3650 by 36.5). By specifying `points=.` you may request the calculations be done at the all the death or censoring time points found in the data set (see section 4.4). This will work fine unless the dataset is quite large (> 500); depending on your environment SAS may not have enough space for the macro's arrays.

For the conditional estimate, the statement would be:

```
* Conditional method;
%survexp(data=jasa, pop=US_WHW, birthdt=birth_dt, firstdt=entry_dt,
    time=fu_time, sex=sex, race=race, method=3,
    points=0 to 3650 by 18.25);
```

The conditional method uses the same time as the Kaplan-Meier, therefore the parameter `time` is required. Using the `points` parameter as stated will generate estimates at every 18.25 days for 10 years. The population selected is one using race, hence the parameter `race` is included. Printing and plotting options are available, as above.

If you want to use a user population, as in the smoking example shown earlier, you must first create the population dataset similar to that described in Section 2.3.2. To use a covariate other than age, sex or calendar year within a user population, you must "borrow" the race variable (set it to your covariate). The SAS code to create the user smoking hazard values for use within `%survexp` or `%lpt` would be as follows. An initial data step that reads the tabular data to create the `smoke` data set is not shown.

```
data lt_user;
    set smoke;   /*  smoking rates stored as one obs. per rate
                         with the variables:
                             sex (m,f)
                             age (49,54,59,64,69,74,79)
                             smoking (a 2 character variable having 15
                                       values corresponding to the 15
                                       possible smoking classes)
                             rate=rate value   */
    keep pop year sex race age q hazard;
    retain oldage;

    pop='smok';        /* pop can be any character value */
    race=smoking;      /* use the variable Race for your covariate */
    year=1970;         /* use an arbitrary decade value */

    if age=49 then oldage=0;
    cage=age;

    do i=oldage to cage;   /*  generate one obs. per age, beginning at 0 */
         age=i;
         q=rate/100000;              /* convert from rate to hazard */
         hazard=-log(1-q)/365.241;
         output;
         end;
    oldage=cage;
```

```
if cage=79 then do;     /*  generate obs up to age 109 */
    do i=80 to 109;
       age=i;
       q=rate/100000;
       hazard=-log(1-q)/365.241;
       output;
       end;
    end;
return;

%survexp(data=jasa, pop=user, birthdt=birth_dt, firstdt=entry_dt,
    sex=sex, method=2, points=0 to 7300 by 182.5, lastdt=p_date,
    race=sm_stat);
```

In the above call to %survexp, the parameter race is pointing to the smoking status variable sm_stat within the input dataset. This would be a two character variable, containing the same codes as were used in the reference population out.smoke further above.

Appendix1: Differences between S and SAS

A few small differences exist in the implementation of the S and SAS functions for these methods. As a result, the computed estimates may differ in the sixth or seventh decimal place. They are documented here to forstall any worries that this might cause.

6.1 Approximations

As documented in section 4.4, the calculations done at each observed death or censoring time may be replaced by an approximate calculation using longer intervals of time. By default, the SAS code uses the approximate calculation with time intervals of 30.5 days and the S code uses the exact calculation. In each this is an option that can be reset by the user.

6.2 Leap year

The S calculations are controlled by the cutpoints attribute of the rate table. The age dimension of this attribute is the rounded value of (0:109) * 365.241 which gives 365, 730, 1096, 1461, ... Depending on the relationship of a subject's birth year to the next leap year, he/she might be 366 days old on their first birthday rather than 365 days old. The error that this introduces is extremely small: a subject might be given 1 extra day

at the age 21 rates and one less at the age 20 rates, for instance, than he would have in a "perfect" computation.

The SAS code uses date subtraction to compute all of its intervals, e.g., 3/1/89 - 3/1/88 to compute the number days spent in an interval from March 1 to March 1. The code automatically creates a subject's birthdates and the anniversary of his/her enrollment date by repeatedly adding +1 to the year. The avoids the leap year problem mentioned above; however, the code fails if a the birth or entry date is Febuary 29, since adding 1 to the year will give an invalid date. To avoid this any birth or entry date of Feb 29 is changed to Feb 28.

6.3 User rate tables

A potential difference between user entered tables and the US population data concerns the special interaction found in the latter between age and calendar year. In an arbitrary rate table containing these variables, we would usually use the (20, 1963) entry to compute a 20 year old subject's one-day hazard on 5/10/63, rather than the (20, 1962) entry as is done in the US tables, based calendar year of that subject's last birthday (8/30/62). In the S implementation, the US and state rate tables have a special flag which signals the "US" behavior to the `survexp` function. This flag also causes automatic interpolation over calendar year. This flag will normally not be present in a user created rate table.

The SAS macro will always use the US behavior.

Appendix2: Corrections to the data

Figure 1 shows an excerpt from the original US data used in `proc survfit`. The US rates for ages > 100 are very unstable. Plots of data for the Minnesota data tables (not shown) reveal the same problem. The published tables from which these data were entered contain both a column of survival probabilities p_i and an integer column L_i giving the expected number still alive out of a cohort of 100,000 subjects

$$L_i = \text{round}(100,000 \times p_i).$$

For convenience, it was this latter number which was transcribed when the Mayo computer tables were first created. The survival probabilities p_i were then recovered as $\hat{p}_i = L_i/L_{i-1}$. For the higher ages, however, L_i may be less than 10, which introduces significant round off error. Even for the lower ages the accuracy will not be the full 5 digits contained in the original data. The actual impact in most studies would be very small, however, since very few person years are contributed by these extreme ages.

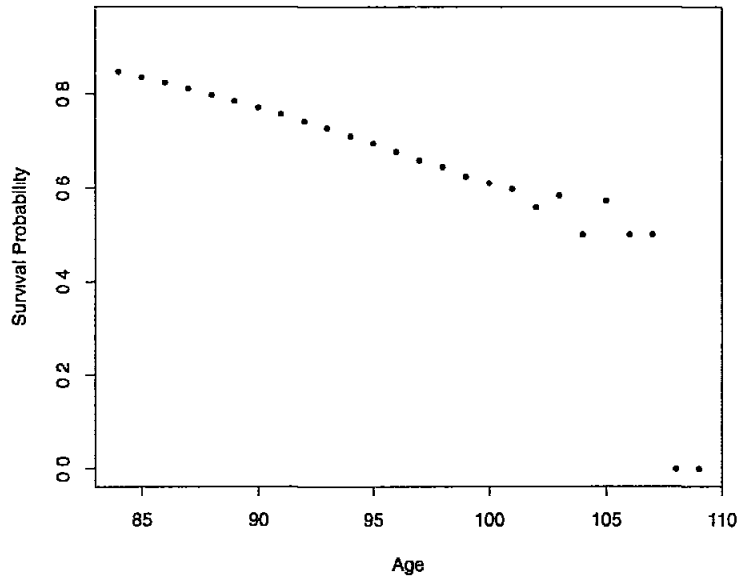

Figure 2: *Survival Probabilities for US white males in 1950*

The raw data for the WNC data is arranged by age, sex and calender year. The years are the decades from 1910 to 1980, and the age groups are 0–6 months, 6 months to 1 year, and then integer years 2, 3, ..., 109. This data was gathered from several different sources; details are found in [2]. Pulling all of this together was an outstanding piece of work, and the issues discussed below are relatively minor in comparison.

Plots of the WNC data are shown in figures 3 to 5 for males, and figures 6 to 8 for females. The hazard over an interval is, by definition, the logarithm of the conditional survival for that interval. Plots are based on the log hazard, which helps to spread out the curves. Each curve is labeled with the second digit of the age. The plots are interesting as a marker of population patterns. There is a general decrease in death rate from 1920 to 1980, with notable exceptions: Figure 3.3 shows the transient increase in hazard for young men aged 21-25 in 1970, and for those aged 25-30 in 1930. The curves for females generally mimic the structure for males, but with fewer unusual features.

The rounding problem found in the US tables is not present here, so no new data had to be re-entered. However, the source for the WNC table was the Minnesota rate table for any years after 1960 (DHEW ceased to publish a WNC table), so the 1960 to 1980 WNC rates were replaced with the new Minnesota rates to preserve consistency.

There is a clear data error in Figure 4.2 for males aged 58 and 59. In response, the two data points have been replaced by a linear interpolation (on the log hazard scale) of

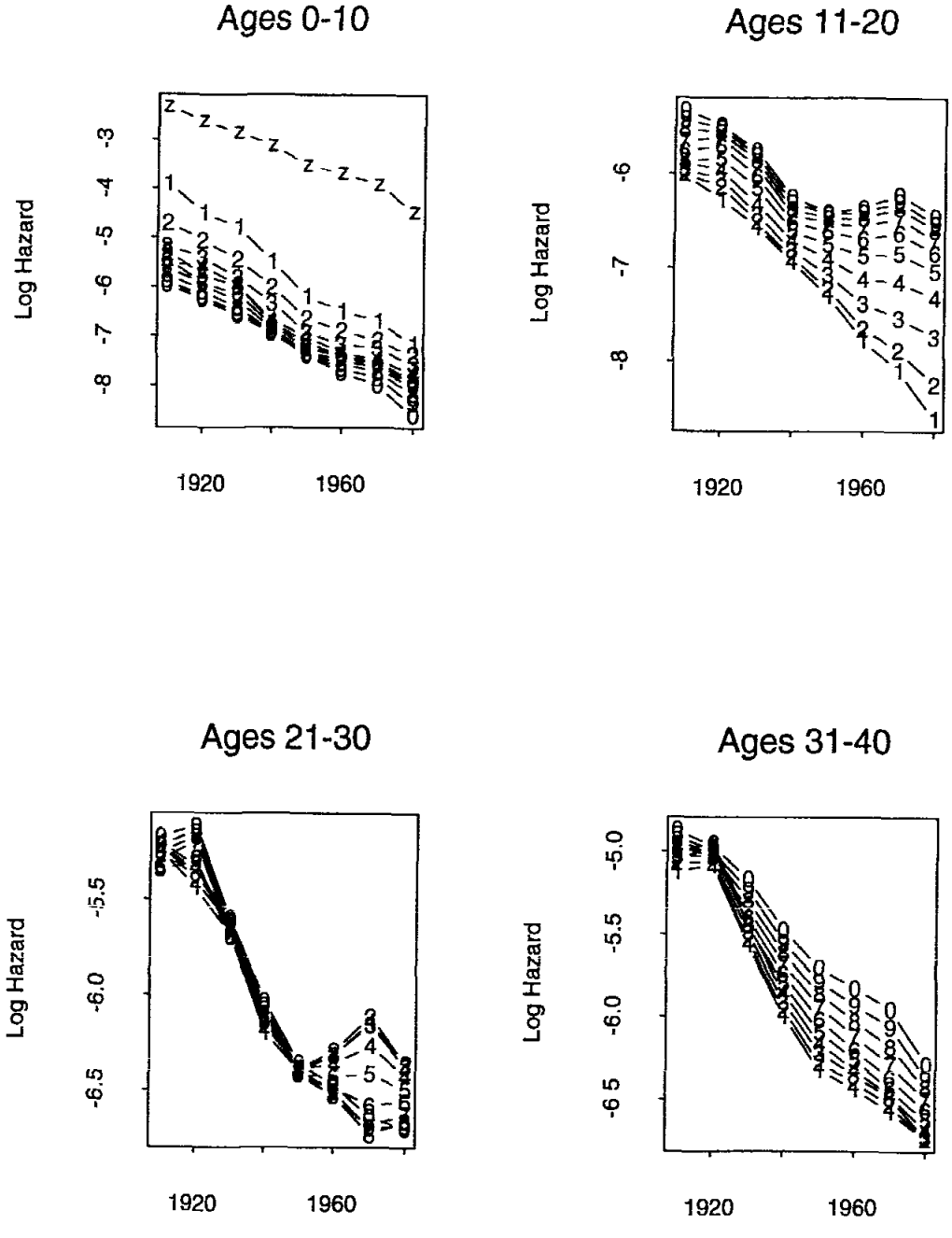

Figure 3 *Log Hazard for West North Central Males Ages 0-40*

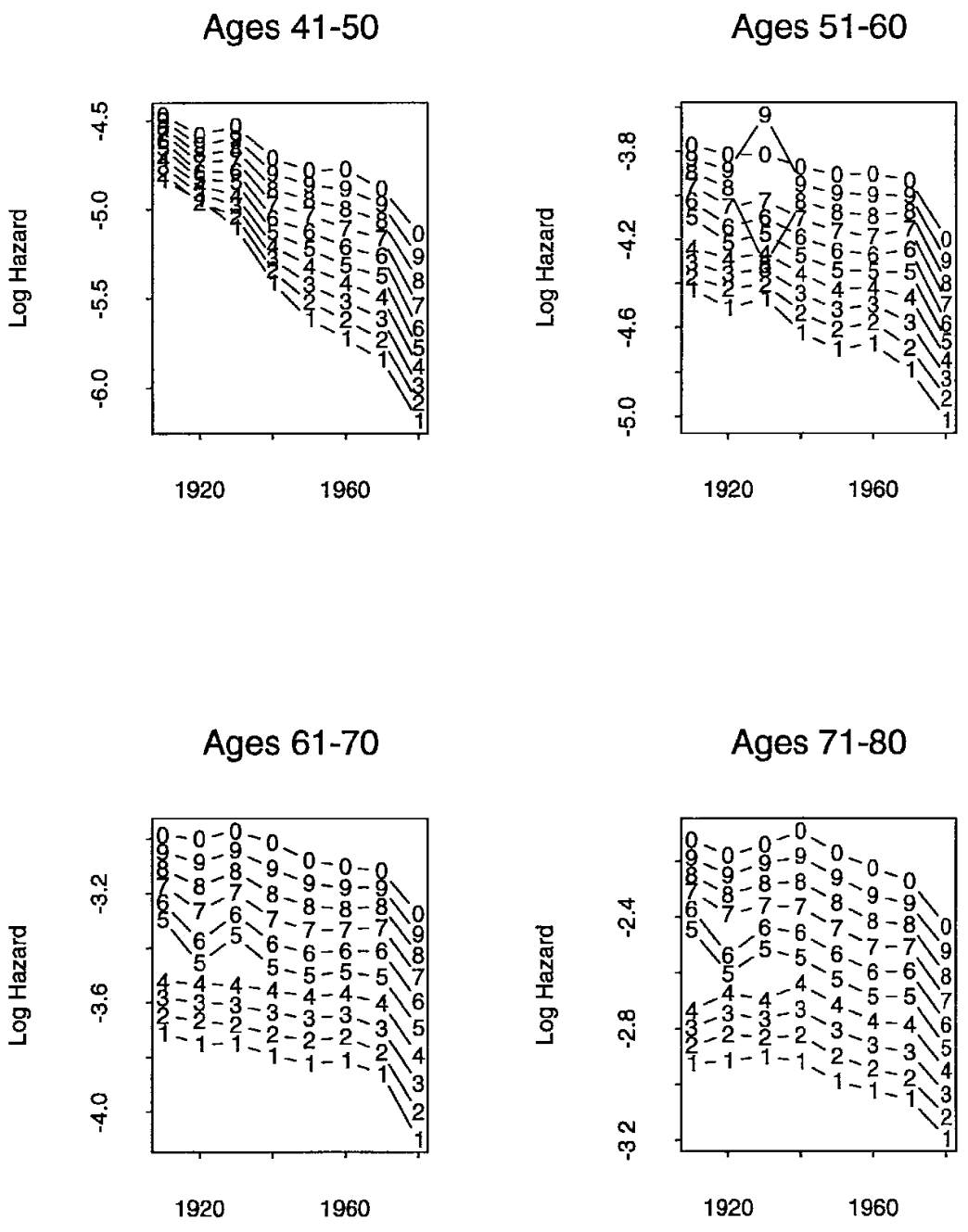

Figure 4: *Log Hazard for West North Central Males Ages 41-80*

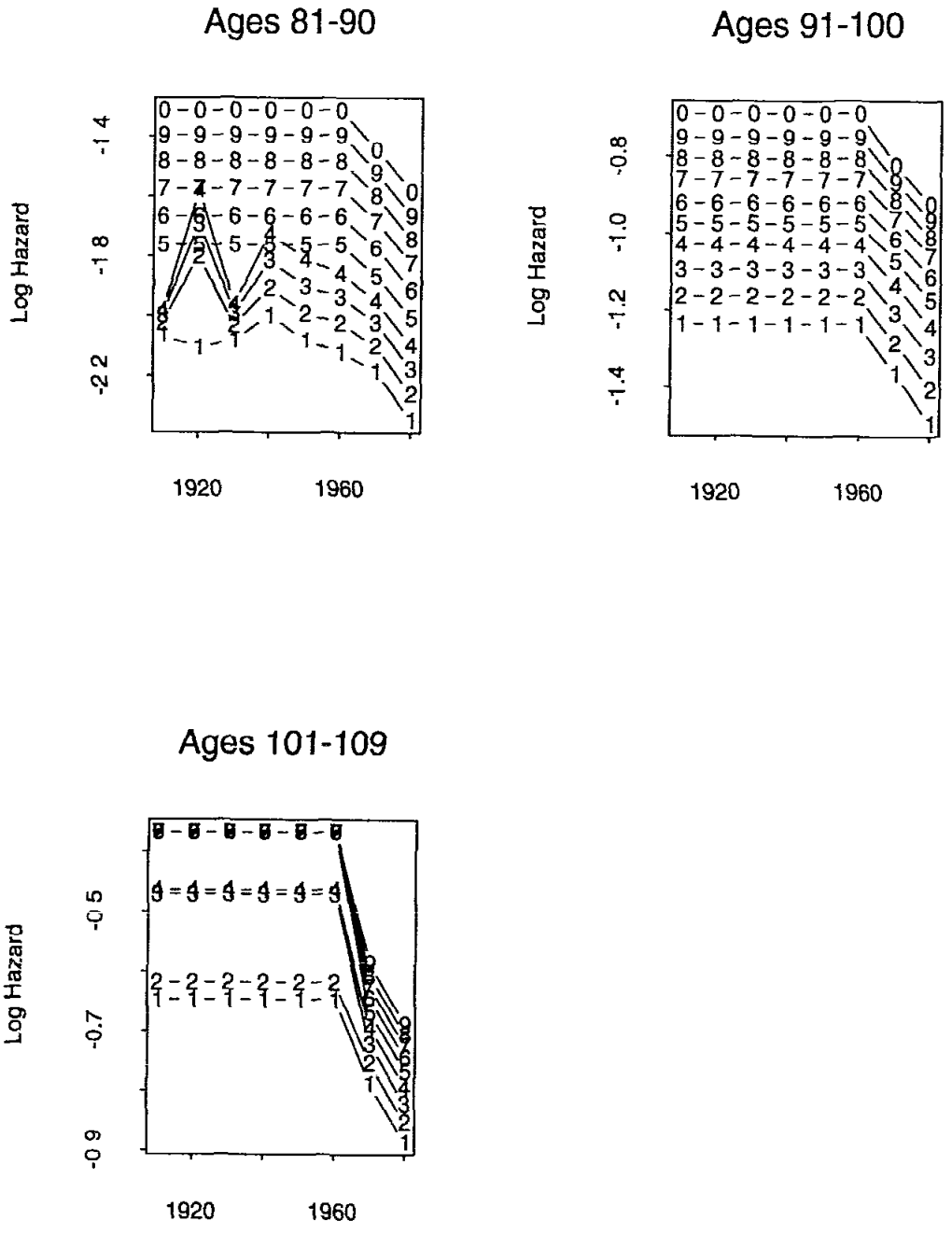

Figure 5: *Log Hazard for West North Central Males Ages 81-109*

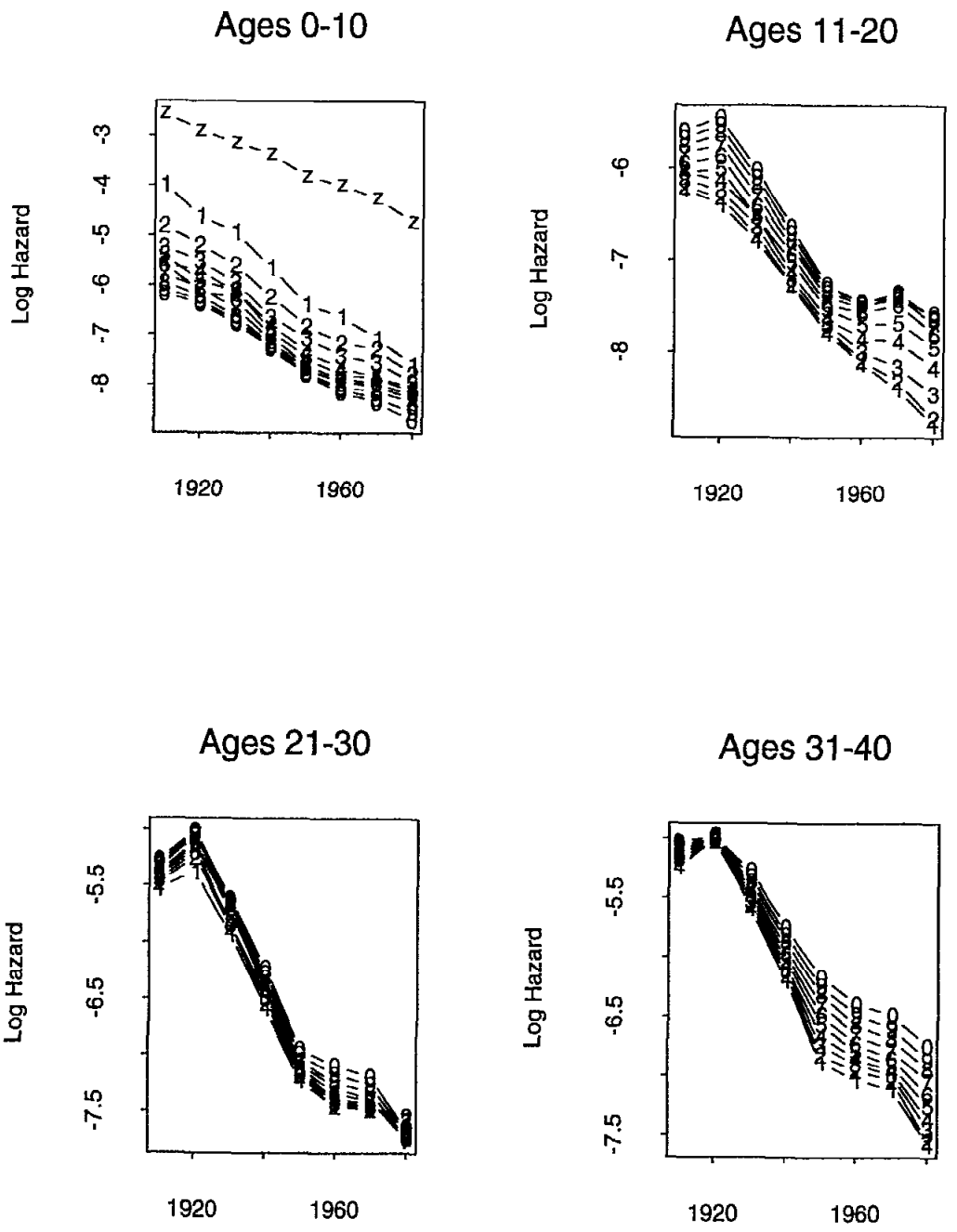

Figure 6: *Log Hazard for West North Central Females Ages 0-40*

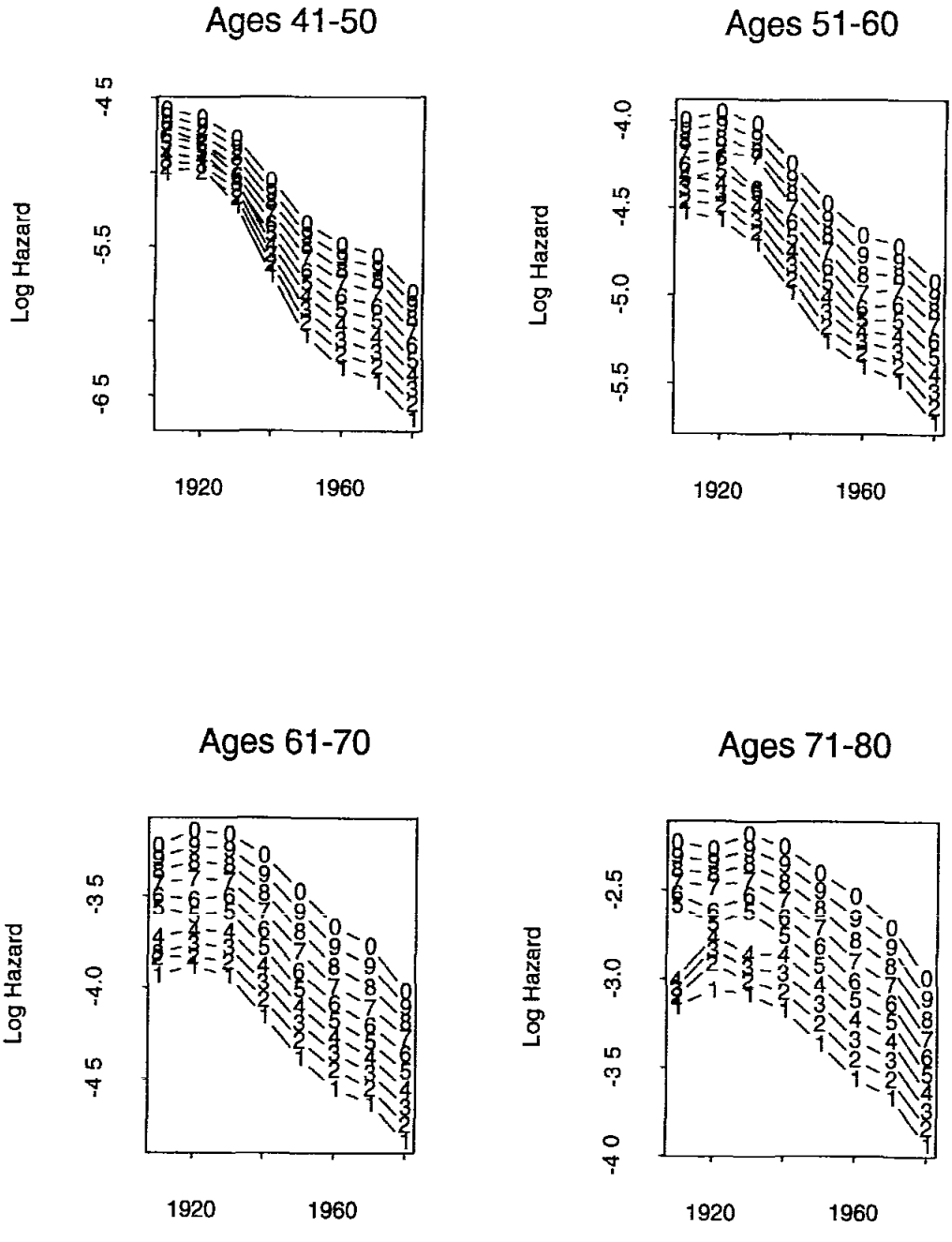

Figure 7: *Log Hazard for West North Central Females Ages 41-80*

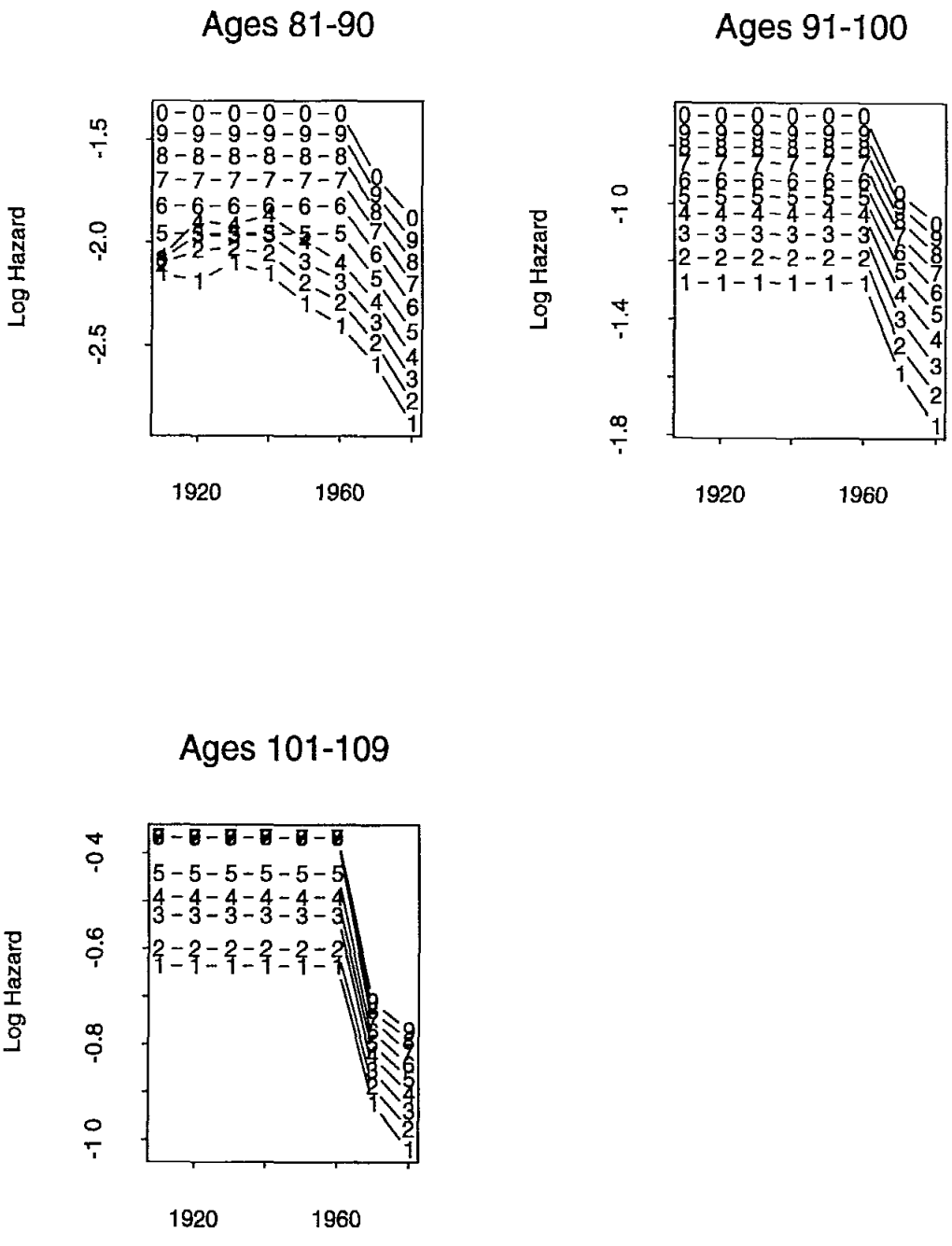

Figure 8: *Log Hazard for West North Central Females Ages 81-109*

the data at ages 57 and 60. This appears consistent with the 1920 and 1940 data which flank the outlying values. Figures 4.3, 4.4, and 5.1 show rather anomalous behavior for some of the 1920 rates, but no action has been taken to correct them. For years prior to 1960, [2] states that the rates for ages 85 and above were created by extrapolation. Figures 5.2, 5.3, 8.2 and 8.3 reveal some possible shortcomings of this action, but again no action has been taken.

For the data used in `survfit`, interpolation between calendar years was done using a seventh degree polynomial, a separate polynomial was fit to each age–sex combination. Figure 9 shows WNC female survival rates along with their interpolated curves for 3 selected ages. The coefficients of the curves are listed in [2]. The curve for age 59 shows an actual *decrease* in survival probability between 1970 and 1975, a result of the convex shape from 1930 to 1960, but all of the curves are unstable in the outer 1/3 of the the interval of calendar years. A further concern with the polynomial method has been that the values for all years are changed whenever a new decade's data becomes available. For these reasons polynomial interpolation has been replaced by linear interpolation in the new S and SAS functions. Because linear interpolation can be done "on the fly" by the functions themselves, it has also allowed us to extend interpolation to other populations besides the WNC.

The final change to the original data concerns the WNC half year probabilities of

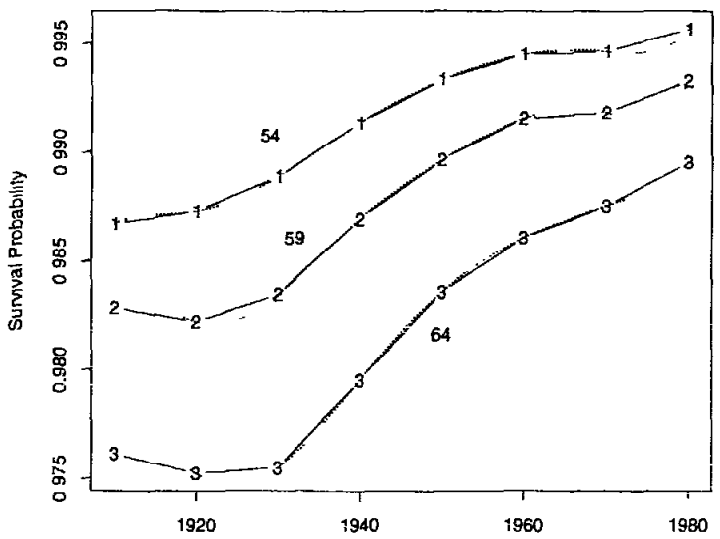

Figure 9 *Survival Probability with Polynomial for West North Central Females Ages 54,59,64*

survival

$$p'_0 \equiv P(\text{survival to age .5})$$
$$p'_{.5} \equiv P(\text{survival from age .5 to age 1}).$$

These numbers were derived from a different source than the whole year probabilities $p_i \equiv P(\text{survival from age } i \text{ to age } i+1)$. As a consequence of using different sources, $p_0 \neq p'_0 p'_5$. Also, data was available for all years and not just the decades. Figure

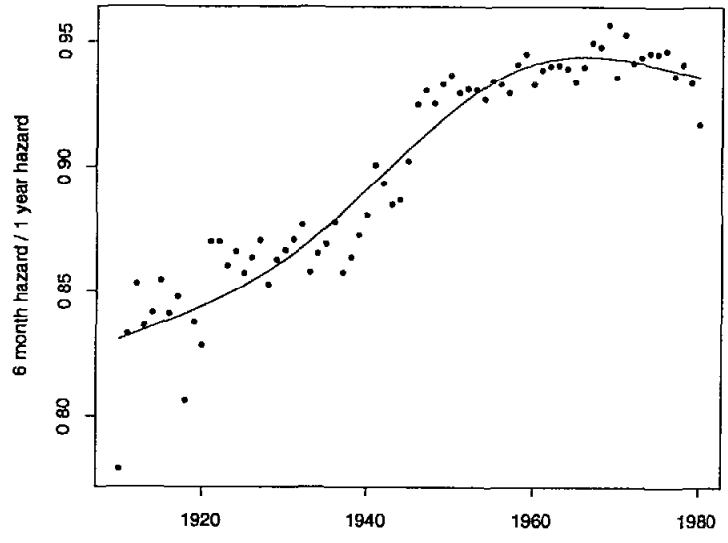

Figure 10: *Ratio of Log Hazard for West North Central Ages 0-6months vs 7-11months*

10 shows a plot of the proportion of the first year's hazard that occurs in the first 6 months, i.e., $\log(p'_0)/\log(p'_0 p'_5)$. These rates are based on small numbers of subjects and are quite variable. New values for the decade years 1910, 1920, ...were obtained by fitting a 3 degree of freedom natural spline to this plot, and partitioning the one year hazard $\log(p_1)$ accordingly. The WNC rate table in S incorporates this half year probability. The half year data was not included in SAS since the macro is less flexible.

References

[1] Andersen, P. and Væth, M. (1989). Simple parametric and nonpararmetric 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 proceedures 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.(198?) *The SURVFIT Procedure.* SUGI Supplemental Library User's Guide, Version 5, SAS Institute Inc., Cary, NC.

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

[14] 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.

[15] *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.