General Models

Terry Therneau

2 Oct 2023

1 Some software notes

The Aalen-Johansen estimate is

$$p(t) = p(0) \prod_{s \le t} T(s)$$

where p(0) is a vector of length k = number of states containing the starting distribution (adds to 1).

At each time s for which a transition occurs, $T(s)$ is the simple transition matrix:

- Row 1 = disposition of all those in state 1 just before time s: (fraction who go to state 1, fraction who go to state 2, ...)
- Row 2 = disposition of all those in state 2 just before time s: (fraction who go to state 1, fraction who go to state 2, ...)
- etc
- Each row sums to 1, all elements are ≥ 0
- If a state j has 0 subjects, set $T_{jj} = 1$. (Makes the computation happy)

If everyone starts in the same state, we normally make that be state 1 and $p(0) = (1, 0, 0, \ldots)$. For the survfit program:

1. The user can set $p(0)$ to anything they want. You then get an estimate of the future probability distribution for a population who all started in that state. Defaut is the observed starting distribution.
2. The user can set start.time to anything they want. This causes only the T matrices at or after that time to be used.

I have found these very useful.

For the coxph program, consider

```
coxph(Surv(time1, time2, state) ~ x1 + x2 + x3, data=zed, id=patient)
```

The default is

1. Each observed transtion type is a separate stratum (separate Cox model)

2. Each model has the same covariates

A more complex form is

```
coxph(list(Surv(time1, time2, state) ~ x1 + x2 + x3,
           1:2 + 1:3 ~ x1*x2 ,
           2:4 + 3:4 ~ x5 + log(x6) / common),
      data= zed, id = patient)
```

1. The first line of the list is the default formula.
2. Line 2 says that two of the transitions also get the interaction x1*x2
3. Line 3 says that 2 other states have the x5 and x6 terms, and that the coefficients for those two are identical.
4. Another option is 'shared', which states that the set of transitions has a shared baseline hazard. This is the opposite of the single state coxph model where strata() forces different hazards. Here the default is to have different ones and the shared statement undoes it.
5. You can have `/common + shared`.
6. Shared (proportional) hazard and identical hazard are not the same.

I am still learning what can be done with these options.

2 Fatty liver disease

Non-alcoholic fatty liver disease (NAFLD) is defined by three criteria: presence of greater than 5% fat in the liver (steatosis), absence of other indications for the steatosis such as excessive alcohol consumption or certain medications, and absence of other liver disease. NAFLD is currently responsible for almost 1/3 of liver transplants and it's impact is growing, it is expected to be a major driver of hepatology practice in the coming decade, driven at least in part by the growing obesity epidemic. The `nafld` data set includes all patients with a NAFLD diagnosis in Olmsted County, Minnesota between 1997 to 2014 along with up to four age and sex matched controls for each case.

We will model the onset of three important components of the metabolic syndrome: diabetes, hypertension, and dyslipidemia, using the model shown below. Subjects have either 0, 1, 2, or all 3 of these metabolic comorbidities.

2.1 Data

The NAFLD data is represented as 3 data sets, `nafld1` has one observation per subject containing basline information (age, sex, etc.), nafld2 has information on repeated laboratory tests, e.g. blood pressure, and nafld3 has information on yes/no endpoints. After the case-control set was assembled, we removed any subjects with less than 7 days of follow-up. These subjects add little information, and it prevents a particular confusion that can occur with a multi-day medical visit where two results from the same encounter have different dates. To protect patient

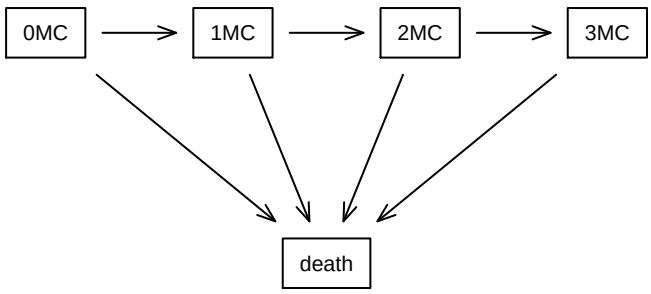

Figure 1: State space figure for NAFLD.

confidentiality all time intervals are in days since the index date; none of the dates from the original data were retained. Subject age is their integer age at the index date, and the subject identifier is an arbitrary integer.

Start by building an analysis data set using `nafld1` and `nafld3`.

```
> ndata <- tmerge(nafld1[,1:8], nafld1, id=id, death= event(futime, status))
> ndata <- tmerge(ndata, subset(nafld3, event=="nafld"), id,
                  nafld= tdc(days))
> ndata <- tmerge(ndata, subset(nafld3, event=="diabetes"), id = id,
                  diabetes = tdc(days), e1= cumevent(days))
> ndata <- tmerge(ndata, subset(nafld3, event=="htn"),  id = id,
                  htn = tdc(days), e2 = cumevent(days))
> ndata <- tmerge(ndata, subset(nafld3, event=="dyslipidemia"), id=id,
                  lipid = tdc(days), e3= cumevent(days))
> ndata <- tmerge(ndata, subset(nafld3, event %in% c("diabetes", "htn",
                                                     "dyslipidemia")),
                  id=id, comorbid= cumevent(days))
> summary(ndata)
Call:
tmerge(data1 = ndata, data2 = subset(nafld3, event %in% c("diabetes",
    "htn", "dyslipidemia")), id = id, comorbid = cumevent(days))

         early late gap within boundary leading trailing tied
death        0    0   0      0        0       0    17549    0
```

```
nafld        0   13   0    318        0    3533        0    0
diabetes  2393    0   0   1058        0       1        0    0
e1        2393    0   0      0     1058       1        0    0
htn       5022    0   0   2045       24       1        5    0
e2        5022    0   0      0     2069       1        5    0
lipid     8663    0   0   1713       82       2        2    0
e3        8663    0   0      0     1795       2        2    0
comorbid 16078    0   0      0     4922       4        7  575
         missid
death         0
nafld         0
diabetes      0
e1            0
htn           0
e2            0
lipid         0
e3            0
comorbid      0
```

A model for the tmerge function is to imagine a drawer with index cards (like an old library card catalog).

- Start with one card per subject, containing a start time, end time, covariate values, and events.
- Make additions one at a time. Each addition has an id,a time point, and a type of addition: a new variable value or new event.
  - The id must already exist.
  - If the timepoint lies within the time range of a card, replace that card with two: (old start, new time), (new time, old end). A new event is added to the first card, a new covariate value to the second. A new variable name or event is propogated to all cards for a subject.
  - If the time point aligns with the start/end of a current card or cards, update them.

The summary function tells us a lot about the creation process. Each addition of a new endpoint or covariate to the data generates one row in the table. Column labels are explained by figure **??**.

- There are 1.7549×10^4 last fu/death additions, which by definition fall at the trailing end of a subject's observation interval: they define the interval.
- There are 13 nafld splits that fall after the end of follow-up ('late'). These are subjects whose first NAFLD fell within a year of the end of their time line, and the one year delay for "confirmed" pushed them over the end. (The time value in the `nafld3` data set is 1 year after the actual notice of NAFLD; no other endpoints have this offset added). The time dependent covariate `nafld` never turns from 0 to 1 for these subjects. (Why were

these subjects not removed earlier by my "at least 7 days of follow-up" rule? They are all controls for someone else and so appear in the data at a younger age than their NAFLD date.)

- 318 subjects have a NAFLD diagnosis between time 0 and last follow-up. These are subjects who were selected as matched controls for another NAFLD case at a particular age, and later were diagnosed with NAFLD themselves.
- 2393 of the diabetes diagnoses are before entry, i.e., these are the prevalent cases. One diagnosis occurred on the day of entry ("leading"), and will not be counted as a post-enrollment endpoint, all the other fall somewhere between study entry and last follow-up.
- Conversely, 5 subjects were diagnosed with hypertension at their final visit ("trailing"). These will be counted as an occurrence of a hypertension event (`e2`), but the time dependent covariate `htn` will never become 1.
- 575 of the total comorbidity counts are tied. These are subjects for whom the first diagnosis of 2 of the 3 conditions happened on the same office visit, the cumulative count will jump by 2. (We will see below that 4 subjects had all 3 on the same day.) Many times ties indicate a data error.

Such a detailed look at data set construction may seem over zealous. Our experience is that issues with covariate and event timing occur in nearly all data sets, large or small. The 13 NAFLD cases "after last follow-up" were for instance both a surprise and a puzzle to us; but we have learned through experience that it is best not to proceed until such puzzles are understood. (This particular one was benign.) If, for instance, some condition is noted at autopsy, do we want the related time dependent covariate to change before or after the death event? Some sort of decision has to be made, and it is better to look and understand than to blindly accept an arbitrary programming default.

2.2 Fits

Create the covariates for current state and the analysis endpoint. It is important that data manipulations like this occur *after* the final `tmerge` call. Successive `tmerge` calls keep track of the time scale, time-dependent and event covariates, passing the information forward from call to call, but this information is lost when the resulting data frame is manipulated. (The loss is intentional: we won't know if one of the tracked variables has changed.)

The `tmerge` call above used the cumevent verb to count comorbidities, and the first line below verifies that no subject had diabetes, for instance, coded more than once. For this analysis we think of the three conditions as one-time outcomes, you can't get diabetes twice. When the outcome data set is the result of electronic capture one could easily have a diabetes code at every visit, in which case the cumulative count of all events would not be the total number of distinct comorbidities. In this particular data set the diabetes codes had already been preprocessed so that the data set contains only the first diabetes diagnosis, and likewise with hypertension and dyslipidemia. (In counterpoint, the nafld3 data set has multiple myocardial infarctions for some subjects, since MI can happen more than once.)

```
Call:
survcheck(formula = Surv(age1, age2, event) ~ nafld + male, data = ndata,
    id = id, istate = cstate)

Unique identifiers       Observations        Transitions
             17549              22683               6186

Transitions table:
       to
from     1mc  2mc  3mc death (censored)
  0mc   1829   70    4   263       5705
  1mc      0 1843   28   243       4567
  2mc      0    0 1048   417       3687
  3mc      0    0    0   441       2220
  death    0    0    0     0          0

Number of subjects with 0, 1, ... transitions to each state:
       count
state       0    1   2   3  4
  1mc   15720 1829   0   0  0
  2mc   15636 1913   0   0  0
  3mc   16469 1080   0   0  0
  death 16185 1364   0   0  0
  (any) 12733 3673 938 183 22
```

This is a rich data set with a large number of transitions: over 1/4 of the participants have at least one event, and there are 22 subjects who transition through all 5 possible states (4 transitions). Subjects do not all enter the study in the same state; about 14% have diabetes at the time of recruitment, for instance. One major difference between current state and outcome is that the current state endures across intervals: it is based on `tdc` variables while the outcome is based on `event` operators. If a subject has time-dependent covariates, there may be intermediate intervals where a covariate changed but an outcome did not occur; current state will endure across intervals but the intermediate outcome will be "censor".

We see a number of subjects who "jump" states, e.g., directly from 0 to 2 comorbidities. This serves to remind us that this is actually a model of time until *detected* comorbidity; which will often have such jumps even if the underlying biology is continuous. The data look like the figure below, where the dotted lines are transformations that we observe, but would not be present if the subjects were monitored continuously. A call to the `survcheck` routine is almost mandatory for a complex setup like this, to ensure that the data set which has been built is what you intended to build.

Calling `survcheck` with ~1 on the right hand side or with the covariates for the model on the right hand side will potentially give different event counts, due to the removal of rows with a missing value. Both can be useful summaries. For a multi-state coxph model neither may be exactly correct, however. If the model contains a covariate which applies only to certain transitions, then events that do not depend on that covariate will be retained, while event

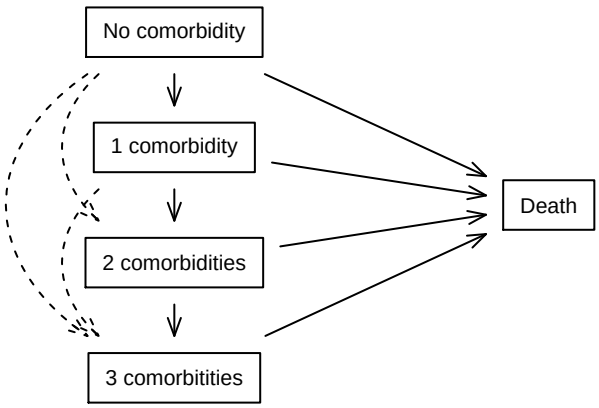

Figure 2: Augmented state space figure.

occurences that do depend on the covariate will be dropped, leading to counts that may be intermediate between the two survcheck outputs.

Since age is the dominant driver of the transitions we have chosen to do the fits directly on age scale rather than model the age effect. We force common coefficients for the transitions from 0 comorbidities to 1, 2 or 3, and for transitions from 1 comorbidity to 2 or 3. This is essentially a model of "any progression" from a given state. We also force the effect of male sex to be the same for any transition to death.

```
> nfit1 <- coxph(list(Surv(age1, age2, event) ~ nafld + male,
                      "0mc":state("1mc", "2mc", "3mc") ~ nafld+ male / common,
                       2:3 + 2:4   ~ nafld + male / common,
                       0:"death" ~ male / common),
                 data=ndata, id=id, istate=cstate)
```

A list has been used as the formula for the `coxph` call. The first element is a standard formula, and will be the default for all of the transitions found in the model. Elements 2–4 of the list are pseudo formulas, which specify a set of states on the left and covariates on the right, along with the optional modifier `/common`. As shown, there are multiple ways to specify a set of transitions either by name or by number, the value 0 is shorthand for "any state". The coefficient matrix reveals that the 1:2, 1:3, and 1:4 transitions all share the same coefficients, as intended.

	NAFLD		Male	
	HR	p	HR	p
0:1-3	2.50	¡0.001	1.20	¡0.001
1:2:3	1.68	¡0.001	1.28	¡0.001
2:3	1.62	¡0.001	1.16	0.0220
0:death	1.88	0.0057	1.39	¡0.001
1:death	1.71	¡0.001	1.39	¡0.001
2:death	1.74	¡0.001	1.39	¡0.001
3:death	1.07	0.4659	1.39	¡0.001

Table 1: Estimated hazard ratio and p-values for the multistate model.

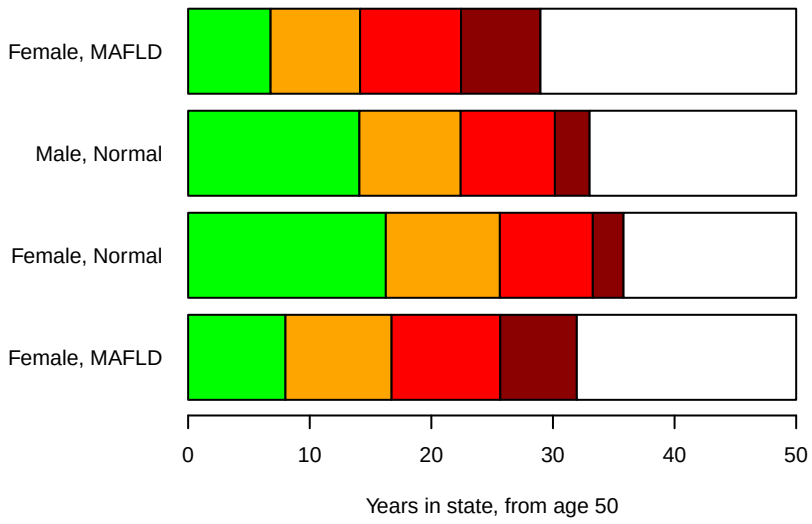

Figure 3: Predicted mean time in state for the MAFLD data. From left to right are 0, 1, 2, 3 cormorbidities, and death (white).