The SAS YOLL macro

Ellen Hertzmark, Molin Wang, Hongyan Huang, Sehee Kim, and Donna Spiegelman

July 17, 2013

Abstract

The SAS YOLL macro uses `PROC PHREG` to compute the time from a specific start time (or age) to an outcome (expected time after the start time to the outcome) or the time from the outcome to a specific time (or age) (expected time lost before the end time). It bootstraps to get the confidence bounds on these values. It computes these values for different levels of an exposure at specified values of the covariates.

Keywords: Years of life lost, expected age at outcome, left truncation, PROC PHREG, bootstrap

Contents

1 Description

Given (possibly left-truncated) data, a time from which you are interested in starting (*TFROM*), a time till which you are interested in 'going' (*TTILL*), the YOLL macro computes the expected time from the *TFROM* to the outcome, and the expected time from the outcome to *TTILL*, as well as confidence intervals for these values. It uses the `BASELINE` statement of `PROC PHREG` to compute the values at the levels of a categorical (indicatorized) *EXPOSURE* and specified levels of the covariates, if any.

2 Invocation and Details

In order to run this macro, your program must know where to look for it. You can tell SAS where to look for macros by using the options

```
options nocenter ps=78 ls=80 replace formdlim='='
mautosource
sasautos=('/usr/local/channing/sasautos',
   '/proj/nhsass/nhsas00/nhstools/sasautos');
```

This will allow you to use all the SAS read macros for the data sets in **/proj/nhsass/nhsas00/nhstools/sasau** as well as other public SAS macros, such as %PM, %INDIC3, %EXCLUDE, %MPHREG, %CALADJ, and %PCTL in **/usr/local/channing/sasautos**.

In the rest of this section, we will list all the input parameters, some of which are required and some of which are optional, but strongly suggested, and some of which are truly optional. NOTE: With this and all other macros, DO NOT include optional parameters in your macro call unless you want to give them non-default values. For example, giving

```
covarforest=,
```

will override the default and cause problems for the running of the macro.

NOTE: if a parameter has a default value, it is given to the right of the '='.

```
PARAMETERS RELATING TO THE DATA AND THE MODEL
=============================================
DATA= The name of the dataset you are using.
      REQUIRED
WHERE=  A subsetting clause, if desired.
        NOTE:  Use 'eq', 'ne', 'gt', etc., rather than '=', '^=', '>'.
        OPTIONAL
TBEG=  Name of the variable for time/age at the beginning of the time interval
```

```
       REQUIRED
TIME=  Name of the variable for survival time within the time interval
       This is the same as for any other proportional hazards model, i.e.
       time from TBEG to the earliest of
           TBEG for the next time interval
           some censoring event.
       One of TIME or TEND is REQUIRED.
       If you give both TIME and TEND, the macro will use TEND.
TEND=  Name of the variable for time/age at the end of the time interval
       This is the earliest of
           TBEG for the next time interval
           some censoring event
       One of TIME or TEND is REQUIRED.
       If you give both TIME and TEND, the macro will use TEND.
EXPOSURE=  The list of exposure variables as indicators
           This can be the macro variable output by %INDIC3, e.g. &exp_ .
           REQUIRED
EVENT=  Name of the censoring/event variable.
        This should be coded so that 0 means censored and 1 means failure.
        REQUIRED
ADJ=  The list of adjusting variables (covariates).
      OPTIONAL
ADJDAT=_basel_  The name of a dataset with values of the covariates
        at which you want to compute the YOLL or time till the EVENT
        This dataset usually has ONE observation.
        An example showing how to make ADJDAT is given in the Examples section.
        REQUIRED if ADJ is not empty
SMALL=.000001  A number to be added to TEND (or TBEG + TIME) when TEND=TBEG, to make TEND gr
       This number should be smaller than the smallest interesting time.
       OPTONAL

PARAMETERS RELATING TO THE TIMES OF INTEREST
============================================
TFROM=  A number (on the same scale as TBEG) at which you want the
        computation to start
        For example, if TBEG is in months and you want your computation
        to start at 50 years of age, TFROM should be given as 600.
        REQUIRED
TTILL=  A number (on the same scale as TBEG) at which you want the
        computation to end
        If the outcome times are bounded (e.g. time to menarche or time to menopause), you n
        to be greater than all reported values of the outcome time.  The specific value you
        the expected survival time, though it will affect the expected time lost.
        In other situations, such as mortality, you may really be interesetd in what happens
        e.g. age 65 and age 75, even though outcomes occur at later ages.  In such as situat
        the end time you are interested in, and the macro will compute the expected survival
        age 75, which is, of course, conditional on survival to age 65.
        REQUIRED

PARAMETERS RELATING TO THE BOOTSTRAP
```

```
====================================
NREPS=500  The number of bootstrap samples
SEED=123456789  The seed for the random number generator to pick the bootstrap samples
```

3 Examples

The example is taken from GUTS. The question is the effect of different types of foods on age at menarche. Since GUTS children were at least 9 at study entry, and some premenarche girls were older than ages at which other girls had attained menarche, the data are left-truncated. *TFROM* is 9 years (=108 months), and *TTILL* is 20 years (=240 months). The output will give the number of months from 108 months of age till the average age at menarche (and its confidence interval). To report the values in a paper, one must do the addition.

3.1 Example 1. Effect of meat consumption on age at menarche

The macro call is

```
%yoll(data=merged, adjdat=cvf,
event=menarche, tbeg=agemo, time=ptime, exposure=&meat_c_,
tfrom=108, ttill=240,
nreps=500,
adj=calor_ &activ_ &inactiv_ &bmi_ &height_ &race_ &bwt_ &mommenar_
&eat_fam_ &man_home_ t9697 t9798 t9899 t9900 );
```

The output is

```
-------------------------------------------------------------------------------------------

Age at menarche
```

Obs	_lineno_	meat_c1	meat_c3	meat_c4	meat_c5	expected survival time between 108 and 240	2.5 pctile survival time	97.5 pctile survival time
2	1	0	0	0	0	28.5964	25.8612	31.499
3	2	1	0	0	0	29.5988	26.2699	32.835
4	3	0	1	0	0	29.0738	26.3873	31.757
5	4	0	0	1	0	28.8672	26.0298	31.322
6	5	0	0	0	1	29.2785	26.4787	31.964

The first several columns give the values for the indicators of meat consumption. Note that the first line is for the reference value, which is level 2. The next 3 columns give the expected months from 9 years (108 months) to menarche and the 95% confidence interval from bootstrapping. The last 3 columns give the expected time from menarche to age 20 (240 months) and the 95% confidence interval from bootstrapping. They are not of great interest in this case, but might be of useful if one were interested in early mortality.

3.2 Example 2. How to make a dataset to use as ADJDAT

Here is the code that made ADJDAT for the age at menarche example.

```
data cvf;
array nums calor_ &activ_ &inactiv_ &bmi_ &height_ &race_ &bwt_ &mommenar_
&man_home_ &eat_fam_  t9697 t9798 t9899 t9900 ;
do over nums;  nums=0;  end;
activ3=1;  inactiv3=1;  bmi3=1;  height3=1;  bwt3=1;
mommenar3=1;  eat_fam3=1;
run;
```

4 WARNINGS

5 Computational methods

5.1 Notation

Let T be the failure time and X be the exposure, and t_1,...,t_m denote the distinct ordered observed failure times, where m is the total number of failures and t_m is the maximum observed failure time. Let t_0 denote the time at which you want your computation to begine (macro parameter *TFROM*). Let t_x denote the time at which you want your computation to stop (macro parameter *TTILL*). Denote the survival function given X $S_X(t) = P[T > t|X]$.

5.2 Calculation of Expected Survival

For simplicity, assume t_m le t_x.

The expected survival =time between t_0 and t_m can be expressed as follow:

$$\begin{aligned}
\mathrm{E}_g[(T-t_0)|X, T>t_0] &= \sum_{k=1}^{m} max(0,(t_k-t_0))\,\mathrm{P}[t_{k-1}<T\le t_k|X,T>t_0] \\
&= \sum_{k=1}^{m} max(0,(t_k-t_0))\,\mathrm{P}[t_{k-1}<T\le t_k|X]/P[T>t_0|X] \\
&= \sum_{k=1}^{m} max(0,(t_k-t_0))\,\{\mathrm{P}[T\le t_k|X]-\mathrm{P}[T\le t_{k-1}|X]\}/S_X(t_0) \\
&= \sum_{k=1}^{m} max(0,(t_k-t_0))\,\{S_X(t_{k-1})-S_X(t_k)\}/S_X(t_0),
\end{aligned}$$

provided that $S_X(t_0) > 0$ for the left truncated survival time.

?? do i have to note that in the case of t(1) gt t(0), the first t(k-1) is t(0), and s(t0)=1. if t(1) eq t(0), then t(1)-t(0)=0 so the summand is 0 ??

If the tail probability of the survival probability is zero, (i.e. $S_X(t_m) = 0$), then $\mathrm{E}_g[T|X, T > t_0]$ is interpretable as expected years of life after t_0.

5.2.1 Dealing with our assumptions

If $t_h <= t_x < t_h + 1,\ldots,t_m$, the calculation can be assumed to have an artificial failure time at t_x with $S_X(t_x)$=$S_X(t_h)$. The sum to this point is the expected time survived from t_0 to t_x, regardless of whether $S_X(t_m) = 0$.

If t_x ¿ t_m, and $S_X(t_m)$ is greater than 0, the expected years of life gained until t_x is between the above sum plus $(t_m - t_0)S_X(t_m)/S_X(t_0)$ and $(t_x - t_0)S_X(t_m)/S_X(t_0)$, but we do not have enough information to know more precisely.

This macro always cuts off the computation at the lower of S_m and S_x. Thus, if $t_x > t_m$, we do not have a true expectation, since the calculation depends on the random variable t_m.

5.3 Calculation of Survival Time Lost before TTILL

NOTE to MOLIN: I want to write the following instead of Sehee's equations below. Note again that, since t_m is random, we can't have a true expectation unless $S_X(t_m)$ is 0 or $t_x < t_m$.

Since we can always write $t_x - t_0 = (t_x - t_k) + (t_k - t_0)$, the expected time lost between t_0 and t_m is

$$[t_m - t_0][S_X(t_0) - S_X(t_m)]/S_X(t_0) - E_g[T|X, T > t_0]$$

provided, of course, that $S_X(t_0) > 0$. MOLIN: the above is not quite right, but i want to give this to you today.

Below are Sehee's equations. again, there is an issue that, since t_m is random, we don't have a real expectation if we cut the calculation at t_m and we have an unknown amount if we don't cut the calculation. Define the time lost before t_m as $U = t_m - T$ for $T < t_m$. Let $u_k = t_m - t_k$. Note that $u_0 > u_1 > \ldots > u_m = 0$.

$$\begin{aligned}
\mathrm{E}_l[U|X, T > t_0] &= \sum_{k=1}^{m} u_k \,\mathrm{P}[u_k \le U < u_{k-1}|X, T > t_0] \\
&= \sum_{k=1}^{m} u_k \,\mathrm{P}[t_m - t_{k-1} < T \le t_m - t_k|X, T > t_0] \\
&= \sum_{k=1}^{m} (t_m - t_k) \left\{S_X(t_m - t_{k-1}) - S_X(t_m - t_k)\right\} / S_X(t_0),
\end{aligned}$$

6 Frequently Asked Questions

6.1 What is the effect of changing TFROM?

6.2 What is the effect of changing TTILL?

7 Description for Methods Section of a Paper

8 Credits

Written by Ellen Hertzmark, Molin Wang, Hongyan Huang, and Donna Spiegelman for the Channing Laboratory. Questions can be directed to Ellen Hertzmark,
stleh@channing.harvard.edu, (617) 432-4597.