

Lexis Expansion- Age-at-Risk Adjustment for Survival Analysis

Lai San Hong, Redsen Limited, Bournemouth, UK

Sarah Lewington, CTSU, Nuffield Dept of Population Health, Oxford University, Oxford, UK

ABSTRACT

Cox proportional hazards model is a widely used method to analyse time-to-event data for survival analysis. In the medical field, age is the most common confounder of any association between risk factor and disease since most risk factors and the risk of disease change with age. Usually, age recorded at the start of the study (age-at-baseline) will be adjusted for in the model to account for the effect of age on risk, but the effect of 'acquired' age during the study (age-at-risk) may be more appropriate because it is age-at-risk that determines current risk rather than the age when the risk factor was measured or determined. Lexis expansion is a method that can facilitate the adjustment of age-at-risk by expanding the observation of each individual into several observations of different age-at-risk bands; age-at-risk can then be adjusted using Cox proportional hazards model. In this paper, Lexis expansion will be explained with examples as well as the implementation of the Cox model using the expanded data.

1. INTRODUCTION

The analysis of time-to-event data is one of the most frequent problems in the field of medicine and biology. For most studies, time since enrolment (or baseline) of the study is used as the origin of time and subjects enter the study at time zero. However, a time scale that is less often considered, namely age 'acquired' during the follow-up of the study (also called age-at-risk or 'attained' age) is often an important determinant of mortality risk. Calendar time is another time scale that could be used for long term follow-up studies. As all these time scales intertwine on the same time axis, a method to separate these scales is beneficial to study the effect of each of them.

Whilst some of the events in survival analysis only occur once (for example, time to death for patients with liver cancer), other events may occur more than once (for example, relapses from remissions for multiple sclerosis patients). We will focus on the analysis of first events (or events that occur only once—eg, death) in long follow-up period.

In this paper, section 2 gives a brief overview of the Cox proportional hazards model and its assumptions; Lexis expansion is introduced in section 3 along with an example; implementation of the expanded data into PROC PHREG is discussed in section 4; and, finally, some conclusions are given in section 5.

2. COX PROPORTIONAL HAZARDS MODEL

The Cox proportional hazards (PH) model [Cox 1972] is a widely used method to analyse time-to-event data. It is based on a hazard function and has the following form:

$$h(t, X) = h_0(t) \exp(\beta'X) \text{ where } X \text{ is a set of covariates.} \quad (1)$$

In this model,

$h(t, X)$ is the hazard for an individual with value X at time t and it is a product of the following two quantities:

$h_0(t)$ is the baseline hazard function and it represents the hazard of an individual at time t with zero values for X (reference group).

$\exp(\beta'X)$ denotes the hazard ratio between group of individuals with value X and the reference group.

Note that the hazard ratio does not depend on time – this is the proportional hazards assumption of the model that the effect of covariates is the same throughout the follow-up time. In other words, the covariate does not have an interaction with time. A covariate is considered time-dependent if it violates this proportional hazards assumption.

3. LEXIS EXPANSION

One approach to handle violation of the proportional hazards assumption by a time-dependent variable like age is to reorganize the structure of the data by splitting the follow-up period of each observation into segments of time intervals. This process of expanding one observation per subject to one observation per follow-up time interval per subject is called Lexis expansion [Lexis 1875, Higgins 2011], which is named after the German statistician and economist, Wilhelm Lexis. Some literature also refers to this process as 'episode splitting' [Jenkins 2008].

For example, consider the following records with one observation per subject:

Table 1.

ID	Born	Entered	Exited	Age-at-entry	Age-at-exit	Time at entry	Time at exit	Follow-up length (years)	Outcome
1	14-Dec-74	30-Jul-04	22-Jan-10	29.63	35.11	0	5.48	5.48	Alive
2	20-Feb-34	03-Sep-84	30-Mar-12	50.54	78.11	0	27.57	27.57	Death
3	17-Aug-51	22-Sep-89	09-Apr-06	38.10	54.64	0	16.54	16.54	Lost of follow-up

In Table 1, entry and exit denote the time when the subject entered and exited the follow-up period. Since each row in this table represents the entire follow-up period of each subject, time at entry is zero for all subjects and time at exit is the same as follow-up length. Also, age-at-entry is the same as age-at-baseline.

To split the above follow-up time into 5-year age-at-risk bands:

Table 2.

ID	Born	Entered	Exited	Age-at-risk band	Age-at-entry	Age-at-exit	Time at entry	Time at exit	Follow-up length (years)	Outcome
1	14-Dec-74	30-Jul-04	14-Dec-04	25-30	29.63	30.00	0	0.37	0.37	Alive
1	14-Dec-74	14-Dec-04	14-Dec-09	30-35	30.00	35.00	0.37	5.37	5.00	Alive
1	14-Dec-74	14-Dec-09	22-Jan-10	35-40	35.00	35.11	5.37	5.48	0.11	Alive
2	20-Feb-34	03-Sep-84	20-Feb-89	50-55	50.54	55.00	0	4.46	4.46	Alive
2	20-Feb-34	20-Feb-89	20-Feb-94	55-60	55.00	60.00	4.46	9.46	5.00	Alive
2	20-Feb-34	20-Feb-94	20-Feb-99	60-65	60.00	65.00	9.46	14.46	5.00	Alive
2	20-Feb-34	20-Feb-99	20-Feb-04	65-70	65.00	70.00	14.46	19.46	5.00	Alive
2	20-Feb-34	20-Feb-04	20-Feb-09	70-75	70.00	75.00	19.46	24.46	5.00	Alive
2	20-Feb-34	20-Feb-09	30-Mar-12	75-80	75.00	78.11	24.46	27.57	3.11	Death
3	17-Aug-51	22-Sep-89	17-Aug-91	35-40	38.10	40.00	0	1.90	1.90	Alive
3	17-Aug-51	17-Aug-91	17-Aug-96	40-45	40.00	45.00	1.90	6.90	5.00	Alive
3	17-Aug-51	17-Aug-96	17-Aug-01	45-50	45.00	50.00	6.90	11.90	5.00	Alive
3	17-Aug-51	17-Aug-01	09-Apr-06	50-55	50.00	54.64	11.90	16.54	4.64	Lost of follow-up

In Table 2, each subject's follow-up time is split into 5 year age-at-risk bands and each row represents the section of the follow-up period when they were in a particular age-at-risk band. Time at entry denotes the amount of time since enrolment to the study when the subject entered the 5-year band; and similarly for time at exit. Outcome denotes the status of the subject at the end of each 5-year band. Therefore, the value of outcome is always alive before the last age-at-risk band.

In this example, the widths of the bands are the same but this is not a requirement. After the expansion, each subject has more than one row of data unless the subject entered and exited the study within the same age-at-risk band. The sum of the follow-up length per subject (person-years) is the same as that before the expansion (Table 1). The number of deaths remains the same after the expansion.

Variables that have different values through follow-up time, i.e. time-varying variable, can be incorporated by specifying different values at different follow-up time interval [Rodríguez 2007]. As mentioned earlier, it is not necessary to have equal width for each interval; indeed, the cut off points can be chosen at the point when the value of variable changes.

LEXIS EXPANSION IN SAS®

To illustrate the implementation of Lexis expansion in SAS, the above example is expanded by using the Lexis macro developed by Carstensen [Carstensen 2007]. First we construct the dataset for Table 1:

```
data table1;
  attrib
    id          label = "ID"
    born        label = "Born"          informat=ddmmyy8.  format=date7.
    entered     label = "Entered"       informat=ddmmyy8.  format=date7.
    exited     label = "Exited"        informat=ddmmyy8.  format=date7.
    age_entry   label = "Age-at-entry"
    age_exit    label = "Age-at-exit"
    fu_yrs      label = "Follow-up length (years)"
    death       label = "Death"
  ;
  input id born entered exited death;
  age_entry = yrdif(born,entered,'act/act');
```

```

    age_exit = yrdif(born,exited,'act/act');
    fu_yrs = yrdif(entered,exited,'act/act');
cards ;
1 14/12/74 30/07/04 22/01/10 0
2 20/02/34 03/09/84 30/03/12 1
3 17/08/51 22/09/89 09/04/06 0
;
run ;

```

To expand Table 1 into 5-year age-at-risk band using the Lexis macro:

```

%Lexis ( data = table1,          /* Original dataset          */
        out = table1_lexis,    /* Name of the output dataset */
        entry = entered,       /* Variable holding the entry date */
        exit = exited,        /* Variable holding the exit date */
        fail = death,          /* Variable holding the exit status */
        origin = born,         /* Origin of the scale where breaks are given */
        left = age_from,       /* Variable for the left endpoint of age at risk band */
        right = age_to,        /* Variable for the right endpoint of age at risk band */
        scale = 365.25,        /* Factor to transform from the scale of entry/exit to
                                the scale where breaks are given */
        breaks = %str( 25,30,35,40,45,50,55,60,65,70,75,80 ) /* Cut off points */
        ) ;

```

The output dataset contains the following variables for each age-at-risk band: the dates of entry and exit; exit status; left and right endpoint of age-at-risk band; follow-up length (risk time in interval) and its natural log value. To obtain age-at-risk band, age-at-entry/exit, and time at entry/exit, some simple calculations are required as shown in the following data step:

```

data table2;
  attrib
    id          label = "ID"
    born        label = "Born"          format=date7.
    entered     label = "Entered"       format=date7.
    exited     label = "Exited"        format=date7.
    age_band    label = "Age-at-risk band" format=$5.
    age_entry   label = "Age-at-entry"
    age_exit    label = "Age-at-exit"
    time_entry  label = "Time at entry"
    time_exit   label = "Time at exit"
    fu_yrs      label = "Follow-up length (years)"
    death       label = "Death"
  ;

  set table1_lexis (rename = (age_entry = age_entry_old risk = fu_yrs)
                    drop = age_exit fu_yrs lrisk);
  time_entry = yrdif(born,entered,'act/act') - age_entry_old;
  time_exit = time_entry + fu_yrs;
  age_entry = yrdif(born,entered,'act/act');
  age_exit = yrdif(born,exited,'act/act');
  age_band = put(age_from,2.) || '-' || put(age_to,2.);
  drop age_entry_old age_from age_to;
run;

```

Stata has a function called *stsplit* [Stata Corp 2013] for this expansion while the Epi package [Carstensen 2013] in R also has the same functionality.

MORE THAN ONE TIME SCALE

If there is another time scale that is of interest, Lexis expansion can easily accommodate it by expanding the data again. Based on the above example which has already been expanded by age-at-risk, the data is further expanded by calendar period (only the first subject is shown here):

Table 3.

ID	Born	Entered	Exited	Age-at-risk band	Period	Age-at-entry	Age at exit	Follow-up length (years)	Outcome
1	14-Dec-74	30-Jul-04	14-Dec-04	25-30	2000-2005	29.63	30.00	0.37	Alive
1	14-Dec-74	14-Dec-04	01-Jan-05	30-35	2000-2005	30.00	30.05	0.04	Alive
1	14-Dec-74	01-Jan-05	14-Dec-09	30-35	2005-2010	30.05	35.00	4.95	Alive
1	14-Dec-74	14-Dec-09	01-Jan-10	35-40	2005-2010	35.00	35.05	0.04	Alive
1	14-Dec-74	01-Jan-10	22-Jan-10	35-40	2010-2015	35.05	35.11	0.06	Alive

Determining which cut off points to use for the expansion depends on the distribution of the event of interest. Some literature suggests having a similar number of events in each band [Alison 1995]. Knowledge about the event as well as the intervention (if there is any) are also of important consideration so that the bands are relevant in clinical terms.

As for the number of time intervals, it is a balance between the number of parameters the model has to estimate and the requirement of the proportional hazards assumption within each time interval. If the data is split into too few time intervals, the hazard ratio of some of the time intervals may not be constant. If too many time intervals are introduced, however, there may not be enough data to estimate the large number of parameters.

4. IMPLEMENTATION IN COX PH MODEL

There are two ways to specify response variable(s) in the model statement from PROC PHREG [SAS 2011]. One way is to specify survival time as the only response variable. Each observation of the dataset represent the entire follow-up period of each subject in the study. To give an example using Table 1:

```
proc phreg data=table1;
  model fu_yrs*death(0)= /* some explanatory variables */;
run;
```

Another way is to use a counting process style of input [SAS 2011]. The counting process method has the same data structure as that of Table 2 where one subject is split into multiple observations and each observation contains information during a semi-closed time interval (t1, t2]. Here, t1 and t2 are the start and end of a time interval when the subject is at risk of the occurrence of the event of interest. Event status at t2 is recorded for each observation as the censor variable specified in the model statement. To implement PROC PHREG for Table 2:

```
proc phreg data=table2;
  model (time_entry,time_exit)*death(0)= /* some explanatory variables */;
run;
```

Without adjusting for the effect of any time-varying or time-dependent variable, the same result will be obtained from the above two PHREG procedures with the same explanatory variables. Although these two PHREG procedures can produce identical results with the inclusion of time-varying or time-dependent variables, more programming statements are needed depending on the data structure. For more examples on this, please refer to example 66.7 of the documentation of PROC PHREG [SAS 2011] or page 153 of Survival Analysis Using SAS [Alison 2010].

ADJUSTING AGE-AT-RISK VIA STRATIFICATION

To adjust for age-at-risk, a stratified Cox PH model can be used when the effect of age-at-risk is not of interest [Prospective Studies Collaboration 2002]. This model takes the form:

$$h_j(t, X) = h_{0j}(t) \exp(\beta'X) \quad (2)$$

where X is a set of covariates and j is the index for age-at-risk band.

This model is very similar to Eq. (1) but each age-at-risk band has its own baseline hazard function. The hazard ratio for the covariate of interest is assumed to be the same for any age-at-risk band and it is constant throughout the follow-up time. The dataset for running this model has to have enough observations for each strata. To specify this model in SAS using Table 2:

```
proc phreg data=table2;
  model (time_entry,time_exit)*death(0)= /* some explanatory variables */;
  strata age_band;
run;
```

Note, stratification on age-at-risk is different to that on age-at-baseline. Running PHREG on the expanded data put subjects who entered the same age-at-risk band to the same strata, whereas running PHREG on the unexpanded data put subjects of the same age-at-baseline band to the same strata. To use the above tables as an example, if we create 5-year age-at-baseline bands from age 25 onwards, all three observations will be in a different strata. If we adjust for age-at-risk, both subject ID 1 and 3 will be in '35-40' strata and both subject ID 2 and 3 will be in '50-55' strata. Although each subject may be included in more than one strata, each subject was counted only once at any moment of time during the entire study period as time at entry is set to time at exit of the preceding interval instead of zero (Table 2).

Although age-at-risk can be adjusted for by stratification, the hazard ratio for its main effect cannot be estimated.

ADJUSTING AGE-AT-RISK AS INTERACTION

One advantage of expanding the data into age-at-risk bands is that the hazard ratio of a covariate can be estimated at a specific age-at-risk band. This is accomplished by incorporating an interaction term between age-at-risk term and the covariate in the data step before PROC PHREG [Prospective Studies Collaboration 2002]. A significance test of the interaction term can indicate whether the assumption of proportional hazards is fulfilled for the covariate – that is, whether the covariate is time-dependent.

Assuming that we have a covariate (expo) for which we would like to estimate its age-at-risk specific hazard ratios and it has value 'Y' for subjects who were exposed and 'N' for subjects who were not exposed, then the data step and model with interaction term as described in the above paragraph are:

```

data interaction;
  set table2;
  age_expo = age_band||'-'||expo;
run;

proc phreg data= interaction;
  class age_expo;
  model (time_entry,time_exit)*death(0)= age_expo /* other explanatory variables */;
run;

```

In this model, a constant hazard ratio of `expo` is assumed within each age-at-risk band, but different bands have different hazard ratios. Importantly, the hazard is not estimated and can vary over the follow-up period.

5. CONCLUSION

This paper describes the use of Lexis expansion to adjust for age-at-risk when the proportional hazards assumption is violated. Although one disadvantage of this Lexis expansion is that the number of observations in the dataset is substantially increased which in turn requires more computing power, the expanded data structure can easily accommodate time-varying covariates, time-dependent variables as well as the effect of other time scale variables. This model does not require a pre-determined distribution for the risk function which gives more flexibility than parametric models. With the computing power currently available, this straightforward data expansion can be easily adopted with existing SAS procedures.

REFERENCES

- Alison, P. (1995). *Survival analysis using SAS®: a practical guide*. SAS Institute.
- Alison, P. (2010). *Survival analysis using SAS®: a practical guide second edition*. SAS Institute.
- Carstensen, B. (2007). *Lexis macro for splitting person-time in SAS*. Available at <http://www.biostat.ku.dk/~bxcl/Lexis/Lexis.sas>.
- Carstensen, B., Plummer, M., Laara, E., and Hills, M. (2013). *Epi: A Package for Statistical Analysis in Epidemiology*. R package version 1.1.49. Available at <http://CRAN.R-project.org/package=Epi>
- Cox, D.R. (1972) *Regression models and life-tables (with discussion)*. J. Roy. Statist. Soc. B 4, 187-220.
- Higgins, C. (2011). *Lecture notes on advanced course in epidemiological analysis 2011. Session 15*. London School of Hygiene & Tropical Medicine.
- Jenkins, S.P. (2008). *Lecture 3: Preparing survival time data for analysis and estimation*. Available at <https://www.iser.essex.ac.uk/resources/survival-analysis-with-stata>.
- Lexis, W. (1875). *Einleitung in die Theorie der Bevölkerungsstatistik*. Trübner, Strassburg.
- Prospective Studies Collaboration. (2002). *Age-specific relevance of usual blood pressure to vascular mortality: a meta-analysis of individual data for one million adults in 61 prospective studies*. Lancet. 360:1903–13.
- Rodríguez, G. (2007). *Lecture notes on generalized linear models*. Available at <http://data.princeton.edu/wws509/notes/>
- SAS Institute Inc. (2011). *SAS/STAT® 9.3 User's Guide*. Cary, NC: SAS Institute Inc.
- StataCorp. (2013). *Stata® Survival analysis and epidemiological tables reference manual. Release 13*. College Station, TX: Stata Press.

ACKNOWLEDGMENTS

Thanks to Dr. Nicolas Christou, Dr. Chihche Lin, Albert Wong, and Dr. Yan Zheng for their remarks and their time on reviewing this paper.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Author	Lai San Hong
Company	Redsen Limited
Address	Bristol & West House, Post Office Road
City / Postcode	Bournemouth, England, BH1 1BL
Email	corinna.hong@gmail.com