# Survival workshop: an annotated guide to the exercises

Contents

[Survival workshop: an annotated guide to the exercises 1](#_Toc87169474)

[Comments 3](#_Toc87169475)

[A100 life table and regression estimates 4](#_Toc87169476)

[A101 life table and regression estimates of conditional survival 6](#_Toc87169477)

[A102 life table and regression estimates (model building) 8](#_Toc87169478)

[A103 conditional net survival (life table estimates) 10](#_Toc87169479)

[A104 standardise regression estimates with meansurvwt 12](#_Toc87169480)

[A105 life table and regression estimates of conditional survival with age standardisaton 13](#_Toc87169481)

[Q101 Kaplan-Meier (life table approach) 15](#_Toc87169482)

[Q102 Kaplan-Meier (life table and product-limit approach) 16](#_Toc87169483)

[Q103 survival and hazard functions 17](#_Toc87169484)

[Q104 K-M survival and statistical tests 18](#_Toc87169485)

[Q110 mortality rates and poisson regression 19](#_Toc87169486)

[Q111 mortality rates and poisson regression with interactions 20](#_Toc87169487)

[Q112 poisson regression with attained age as time scale 21](#_Toc87169488)

[Q120 Cox regression compared to poisson regression 22](#_Toc87169489)

[Q121 Cox regression compared to poisson regression, interactions 23](#_Toc87169490)

[Q122 Cox regression for cause-specific mortality 24](#_Toc87169491)

[Q130 Understanding splines 25](#_Toc87169492)

[Q131 Model cause-specific mortality using flexible parametric models 27](#_Toc87169493)

[Q132 Extending exercise 131 time-dependent effects in flexible parametric models 29](#_Toc87169494)

[Q133 Extending exercise 131 alternative link functions 31](#_Toc87169495)

[Q140 Probability of death in a competing risks framework (cause-specific survival) 32](#_Toc87169496)

[Q201 Life table estimates of relative survival 35](#_Toc87169497)

[Q202 Life table estimates of cause-specific survival 36](#_Toc87169498)

[Q203 Period and complete analysis I 37](#_Toc87169499)

[Q204 Period and cohort analysis II 38](#_Toc87169500)

[Q210 Modelling excess mortality using Poisson regression I 39](#_Toc87169501)

[Q211 Modelling excess mortality using Poisson regression II 40](#_Toc87169502)

[Q230 Modelling excess mortality using flexible parametric models I 41](#_Toc87169503)

[Q231 Modelling excess mortality using flexible parametric models II 42](#_Toc87169504)

[Q232 Modelling excess mortality using flexible parametric models III 44](#_Toc87169505)

[Q240 Age standardization life table methods I 46](#_Toc87169506)

[Q241 Age standardization life table methods II 47](#_Toc87169507)

[Q242 Age standardization using flexible parametric models. 48](#_Toc87169508)

[Q243 Age standardization life table methods III (ICCS weights) 49](#_Toc87169509)

[Q250 Probability of death in a competing risks framework (life table methods) 50](#_Toc87169510)

[Q251 Probability of death in a competing risks framework (regression methods) 51](#_Toc87169511)

[Q261 Cure models 52](#_Toc87169512)

[Q282 Avoidable deaths 53](#_Toc87169513)

[Q284 Residual expectation of life and years of life lost 54](#_Toc87169514)

## Comments

**Introduction and overview**Paul Dickman and Paul Lambert have led several week-long training sessions in modern methods in population-based cancer survival analysis. In the workshop sessions the attendees were offered a series of Stata programs to illustrate the points that they had raised in their lecture presentations. These programs used publicly-available data, so attendees did not need to bring their own data. In preparation for a pre-conference workshop along the same lines, held in Albuqueque, NM in 2017, I offered to write sas equivalents for as many of these Stata exercises as was possible at the time. Since then, a few more have been added, and they have all been reviewed to make sure that they still conform to the needs of the sas macros that they invoke. This documentation is designed to make the resulting programs useful as reference material for the concepts and methods that they illustrate. Several programs have been added to illustrate points that have been brought up since then. Those programs are numbered in the series A100 and following. The original exercise programs are numbered Q101 and following.

**Instructions**The data, macro and exercise folders are defined under a common folder root. The location of the folder root is set as a macro string in the program ‘00 include macros.sas’. This program should be run first in any sas session before running any of the exercises. Once that program has been run any one of the desired exercises can be loaded and run. Each exercise is designed to be run from beginning to end, with some later parts depending on earlier results. Some of the exercises – especially the longer ones - have section markers that can be used for navigation in the corresponding annotations.   
  
In general, data is loaded, sometimes with restrictions and temporary variables computed at the top of the program. If there is a focus on relative survival (either in a regression framework, or a life table approach), a population mortality table is also referred to directly from the data library, but not loaded into the work library. All necessary formats for the data files are available in format libraries in the data library. Since some of the format names are the same for different datasets (for example, stage has different values for the melanoma and for the colon dataset), dataset-specific format libraries are assigned as needed as a first step in each exercise.

## A100 life table and regression estimates

**Data**  melanoma

**Procs**  sgplot

**Macros** %rel\_surv %stset %stpm2 %predict

**Comments** This is based on an example provided by Paul Lambert to show the equivalence between life table estimates and those available from excess hazard regression. See A102 for a further examination of the relationship between life table and regression estimates.

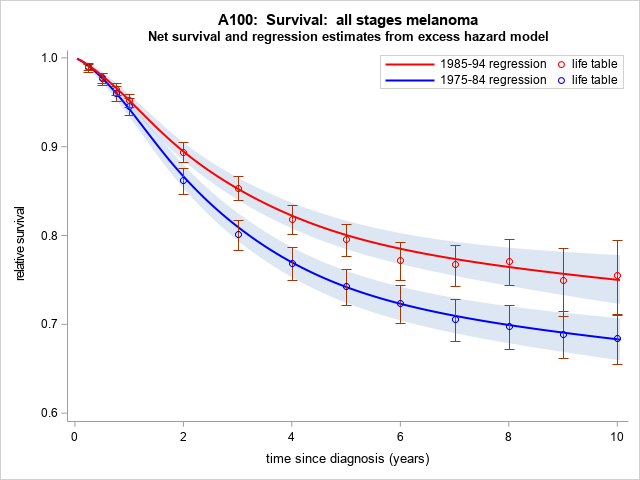
(a) The melanoma data is processed by *%rel\_surv* to compute survival estimates for the two time periods. No results are reported. Estimates are retained in the default (‘*grouped*’) dataset. Note that these estimates can be considered to be age standardized to the age distribution of the cases in the two strata. See exercise [A104](#_Exercise_A104_standardise) and [A105](#_Exercise_A105_life) for the analysis using external standardization to an ICSS standard.

(b) Append mortality rates from the population life tables, matched on sex and age attained at death/censoring, and on year attained at death/censoring. Follow-up is truncated at 10 years, age is truncated at 99 years (mortality rates for older ages are held at that of 99 years), and attained year is held at a maximum of 2000 (the most recent year in the life table supplied).

(c) *%stset* is used to build the analytic dataset. Note the use of the ‘*noprin*t’ option. *%Rcsgen* is used to create 3 splines to describe the age relation. The knot points for the splines are determined from the distribution of ages of patients who have died (‘*if1 = \_death\_ = 1’*), which will insure approximately equal numbers of deaths between knot points. This may improve the fit of the model. The relative survival regression model is fit with the time period variable and all age splines as *TVC* variables. A further 3 splines are used to describe the interaction between each tvc covariate and follow-up time.

(d) *%range* creates a temporary time variable to speed up processing, especially since confidence intervals are request. *%Predict* is used with the *meansurv* measure to obtain relative survival estimates for each time period. The *meansurv* measure obtains the average of the survival curves for each distinct covariate pattern, weighted by the observed distribution of that covariate pattern in the data. Each of the two time periods is estimated separately, by selecting the data for that time period with the ‘*ifp*’ parameter. This results in estimates which are internally age-standardised, the weights being specific to the time period strata

(e) the Pohar-Perme net survival estimates from step (a) and the regression estimates from (d) are plotted on the same time scale. The estimates from the later time period are somewhat less precise that those from the earlier, especially at 10 years. The 10-year estimate for the earliest time period is based on cases diagnosed in the whole period from 1975 to 1984, since there is follow-up to the end of 1995 for this dataset. For the later time period, the 10-year estimate can only be estimated from the 1985 cases. There is insufficient follow-up time for later cases. The resulting plot is below.



## A101 life table and regression estimates of conditional survival

**Data**  melanoma

**Procs**  sgplot

**Macros** %rel\_surv %stset %stpm2 %predict

**Comments** This is based on an example ([A100](#_A100_life_table)) provided by Paul Lambert to show the equivalence between life table estimates and those available from excess hazard regression. It has been expanded here to show how to compute conditional survival estimates, a feature that was not present in the version of *%predict* that was presented in Albuquerque in 2017. See A103 for an example that combines conditional survival probabilities and external age standardization.

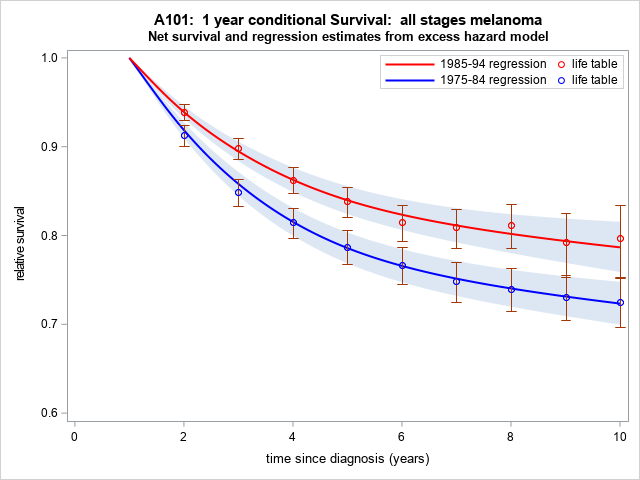
(a) Create an entry time as the date one year after the date of diagnosis. This is used in the %rel\_surv call to restrict analysis to patients who have survived the first year after diagnosis. This is known as a ‘late entry’ analysis (this is similar to period analysis. See A10x for the corresponding code for using *time* since diagnosis to define the entry time. Annual intervals are chosen, as the excess hazard does not change rapidly after the first year of follow-up. The melanoma data is processed by *%rel\_surv* to compute net survival estimates for the two time periods. No results are reported. Estimates are retained in the default dataset ‘grouped’.

(b) Append mortality rates from the population life tables, matched on sex and age attained at death/censoring, and on year attained at death/censoring. Follow-up is truncated at 10 years, age is truncated at 99 years (mortality rates for older ages are held at that of 99 years), and attained year is held at a maximum of 2000 (the most recent year in the life table supplied).

(c) *%stset* is used to build the analytic dataset. Note the use of the ‘*noprin*t’ option. *%Rcsgen* is used to create 3 splines to describe the age relation. The knot points for the splines are determined from the distribution of ages of patients who have died (‘*if1 = \_death\_ = 1’*), which will insure approximately equal numbers of deaths between each knot point. This may improve the fit of the model. However, it does not appear to have had any impact on the likelihood. The relative survival regression model is fit with the time period variable and all age splines as *TVC* variables. A further 3 splines are used to describe the interaction between each covariate and follow-up time.

(d) *%predict* is used with the *meansurv* measure to obtain relative survival estimates for each time period. Note the use of the temporary time variable created with *%range*. The *‘tcond’* parameter specifies that conditional survival estimates are requested, the condition being that patients must first survive one year from diagnosis. Each time period is estimated separately, by selecting the data for that time period with the ‘*ifp*’ parameter. This results in estimates which are internally age-standardised, the weights being specific to the time period strata.

(e) the Pohar-Perme net survival estimates from step (e) and the regression estimates from (d) are combined and plotted on the same time scale.



## A102 life table and regression estimates (model building)

**Data**  melanoma

**Procs**  sgplot

**Macros** %rel\_surv %stset %stpm2 %predict

**Comments** Further investigation of the differences between life table and regression estimates of relative survival, and how modeling can be used for consistent estimation. A series of models are fit, followed by plots of the survival and hazard estimates. All output is saved in an excel spreadsheet (‘*life table vs regression.xlsx*’ in the ‘reports’ folder), which has been manipulated to bring the plots in adjacent tabs. A summary table of likelihood ratio tests has been added, based on the fit statistics of each modelling step. The ODS statements have been commented out in this exercise.

(a) As in exercises [A100](#_A100_life_table) and [A101](#_A101_life_table), use *%rel\_surv* to compute life table estimates of net survival. Interval-specific excess hazard estimates are also saved in the ‘grouped’ file and will be needed in later steps. No stratification is specified.

(b) Append population mortality rates

(c) Use *%stset* to prepare the analytic dataset, and use *%rcsgen* to generate 3 splines to describe the age relation. Also set up the ODS commands to capture all printed and plotted output into separate spreadsheet tabs.

(d) Each candidate model is fit, followed by plots of the predicted survival and hazard functions. These plots are combined with the corresponding estimates from the life table analysis. The *%smooth* macro is used to compute a hazard function based on the computed survival function. This is not necessary for the ‘base’ model, as %predict could have been used to compute the hazard function. However, for the subsequent models, where the *meansurv* option is used to compute the averaged survival function, there is no corresponding *%predict* measure that can be used to compute an ‘averaged’ hazard function in the same way. So *%smooth* is used as an alternative. In order to get a good approximation to the survival function (especially in the first year of follow-up), the survival function is estimated at 100 intervals per year. The hazard rate (*lambda*, from *%smooth*) and the excess hazard from *%rel\_surv* are scaled to a rate per 1,000 person-years.

The models to be fit are  
base model no covariates 6 parameters  
age add the age splines as covariates 9 parameters  
Age sex add sex to the previous model 10 parameters  
age (TVC) sex add age as a TVC to the ‘age sex’ model 19 parameters  
age sex (TVC) add sex as a TVC to the ‘age sex’ model 13 parameters  
both TVC both age and sex are added as TVC variates 22 parameters

The base model does not recover the survival estimates well. The reason can be seen by examining the hazard plots. The hazards are well captured in the first year, but drift off subsequently. This is likely a result of the change in the age distribution of the surviving patients, a change which is accounted for in the life table estimates, but not in the base model. Adding age to the model helps greatly. Adding sex as well improves the fit (according to the LR test), but is not apparent in the survival plot. Further detailed modelling with TVC variables has an impact, but hardly seems worth the effort.

**ODS trick:** note the use of the *sheet interval = ‘now’* and *sheet interval = ‘none’* to control the way the spreadsheet tabs are created, and to keep printed output on one sheet and combined plotted output on another.

Likelihood ratio tests for the above models

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **model** | **-2 Log Likelihood** | **df** | **chi sq** | **p-value** | **test against** |
| **base** | 17176 | 6 |  |  |  |
| **age** | 17082 | 9 | 94 | 0.0000 | base |
| **age sex** | 17003 | 10 | 79 | 0.0000 | age only |
| **TVC models** |  |  |  |  |  |
| **age sex(tvc)** | 16996 | 13 | 7 | 0.0677 | age and sex |
| **age(tvc) sex** | 16987 | 19 | 16 | 0.0610 | age and sex |
| **both tvc** | 16979 | 22 | 24 | 0.0193 | age and sex |

## A103 conditional net survival (life table estimates)

**Data**  colon

**Procs**

**Macros** %rel\_surv

**Comments** This is an example drawn from on-line tutorials provided by Paul Dickman (see *https://www.pauldickman.com/software/stata/* for full list of tutorials). Estimating survival conditional on an initial period of survival can be done by simple division of survival estimates, but confidence intervals do not result. This exercise shows how to set up data for conditional estimation using life tables. Both cohort and period-based approaches to estimation are given, with the code for setting up the entry time (or date) for the conditional analysis.

The method for the conditional computation extends to age-standardised estimates with the inclusion of the ‘*stnd =*’ and ‘*weight\_lib’* directives. (used in exercises [Q[24](#_Exercise_Q204_)0](#_Exercise_Q240_) and [Q241](#_Exercise_Q241_)). An ‘ad hoc’ estimate based on the ratio of unconditional age-standardised is not correct but might be reasonably close.

The first section describes the cohort approach.

*%rel\_surv* is used to produce life table estimates of relative survival. The interest is in computing the relative survival at 5 years post diagnosis, conditional on having already survived 1 year. Considering the relative survival estimates as probabilities, this estimate can be obtained by dividing the 5-yr estimate by the 1-yr estimate. Both methods of describing the risk times are given.

extract the 1-yr and 5-yr estimates from the last *%rel\_surv* run, compute and print the 5-yr conditional estimate. The Pohar Net Survival estimates are used here, but the Ederer 2 estimates (*cr*) are available in the same file.

using the ‘*entry =* ‘ directive in *%rel\_surv*, it is possible to restrict the analysis to the patient experience in the intervals after the first year of follow-up. This can be described as the survival experience of patients who have already survived the first year after diagnosis. This is done using dates, set ‘*entry =’* to a variable holding the diagnosis date + 365.24. Alternatively, on the follow-up (duration) scale set ‘*entry =*’ to point to a variable holding the time after which patients come into risk. Since follow-up time is expressed as days in this example, the variable defined is globally 365.24.

In both cases, *%rel\_surv*, uses the default scale factor of 365.24 to convert to the scale of years after the date of diagnosis.

The two alternate versions of *%rel\_surv* are called. For the version that specifies survival on the time scale, the year of diagnosis must be specified with the ‘*yydx =*’, *‘exit =* ‘ should point to the total length of survival. Since the mortality risk does not change rapidly after the first year, annual intervals are chosen. Only the estimate for the 5th year after diagnosis is displayed. If all intervals were to be displayed (or viewed in the ‘*grouped’* file), the first interval reported is for the period 1-2 years after diagnosis. However, since the *‘intervals =* ‘ directive specifies intervals starting at diagnosis, the Pohar weights are computed correctly.

The second section goes through the same steps, but uses the period approach to estimation, only considering patient experience in the latest years available in the dataset: 1993 – 1995. Both the unconditional and conditional analyses require the use of the ‘*entry =*’ directive.

For the unconditional estimates, the entry date is the start of the period window (1 Jan 1993). That unconditional entry time is either 0 (for cases diagnosed within the period window) or the number of days between the start of the period window, and the date of diagnosis.

For the conditional estimates, the time (or date) of entry into the period risk window must be adjusted to allow for an initial survival interval of 1 year. Patients diagnosed within the period window, and for those diagnosed within the year prior to the period window come into risk at a date 1 year after their date of diagnosis. Patients diagnosed in the final year of the period window will have an adjusted entry date that is *after* their exit date, and will be excluded by *%rel\_surv*. Patients diagnosed more than one year before the start of the period window come into risk at the start of the period window.

For the corresponding entry points on the duration scale, the same considerations must be interpreted. Patients diagnosed within the period window, or in the previous year, enter at day 365.24. Patients diagnosed more than one year prior to the start of the period window will come into risk at the start of the period window (on the duration scale) without modification.

## A104 standardise regression estimates with meansurvwt

**Data**  colon (stage 1)

**Procs**  sgplot freq sql means

**Macros** %stset %stpm2 %rcsgen %range %predict

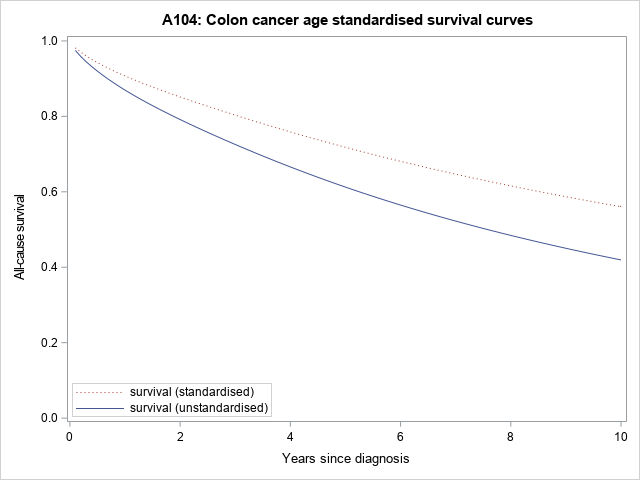
**Comments** This exercise is based on one of Paul Dickman’s online tutorials (see: <https://www.pauldickman.com/software/stata/>) on the topic of age-standardisation of regression estimates using the stpm2 command in Stata. The goal of age standardisation is to estimate the measure (net survival) that would have been observed in the study population, if that population had the same age structure as the reference population. The ICSS weights used in age standardization in *%rel\_surv* are used to compute a weighted average of stratum-specific survival estimates. The corresponding process at the individual level, is to assign a weight to patient records, such that the age distribution of the (weighted) cases is identical to the ICSS reference age distribution.

(a) compute the ICSS age groups using the relevant C-SPAN format, and tabulate the distribution using *proc freq*. Save the observed proportions (as percentages) for later steps.

(b) compute the relevant individual-level weights by scaling the observed age-specific proportions, and assign the weights to the cases selected for analysis

(c) prepare the data for modeling with *%stset* and fit a regression model with %*stpm2* (note that this is not an excess hazard model, so the survival proportions are cannot be interpreted as relative survival estimates). Create a temporary time variable with *%range* and use *%predict* to estimate several survival curves. The first is the population-weighted curve, representing the average across the observed age distribution, using the ‘*meansurv*’ measure. The second call to *%predict* applies the individual-level weights with the ‘*meansurvwt =* ‘ directive.

(d) compare the two survival curves. The resulting plot is copied here



(e) finally, estimate the weighted and unweighted versions of 5-year survival. The population-weighted estimate varies across patients, but mean is the same as what was estimated with *meansurv* in (c) Multiplying each survival estimate by the corresponding individual-level weight produces a survival estimate whose average is identical to the standardized estimate.

## A105 life table and regression estimates of conditional survival with age standardisaton

**Data**  melanoma

**Procs**  freq sql sgplot

**Macros** %rel\_surv %stset %stpm2 %rcsgen %predict %range

**Comments** This exercise takes [A101](#_A101_life_table) (conditional survival estimates with life table and regression estimates) and adds age standardization, using methods presented in [A104](#_A104_standardise_regression).

(a) prepare the data for conditional analysis using *%rel\_surv* by creating an entry time (in months) representing the initial survival period, after which patients come into risk observation. *%rel\_surv* is run, requesting age-standardisation to the ICSS standard 2 with the ‘*stnd = icss(2)’* directive. Display of the age-specific crude estimates (by stratifier) is suppressed by default when age standardization has been requested. The listing that is presented is the age-standardised Pohar net survival estimates. The default destination of the age standardized estimates is over-ridden with the ‘*std\_estimates =*’ directive.

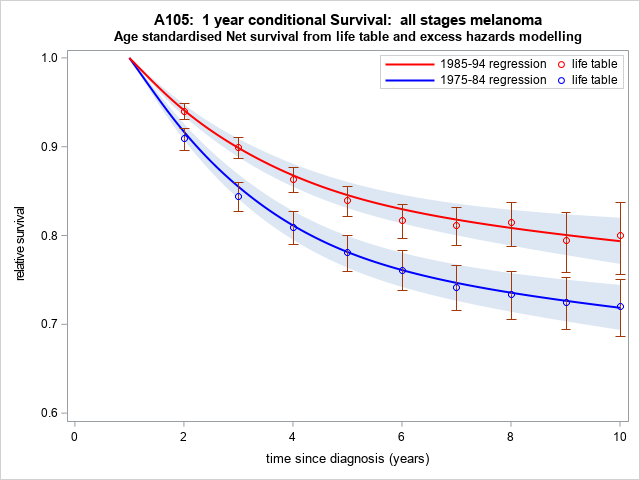
(b) this step appends the population mortality rates, for the excess hazard regression model to be fitted, but also computes and applies the individual-level weights for age standardization. Survival is also truncated at 10 years of follow-up and the status variable is updated accordingly.

(c) *%stset* is used to build the analytic file for analysis, *%rcsgen* is used to create age splines, and an excess hazard model is fit with *%stpm2*. Note that both age and time period of diagnosis are designated as Time-Varying Covariates, with 3 splines to be used to describe the relation between each variable and follow-up time. This is a non-proportional excess hazards model, estimating 22 parameters.

(d) *%range* is used to define a temporary time variable, which will reduce the amount of processing time needed to compute the estimates of weighted-average survival curves (‘*meansurv*’ and ‘*meansurvwt =*’ directives) and confidence intervals with *%predict*. Note that conditional estimates have been requested with patients only coming into risk one year post-diagnosis (‘*tcond = 1’*). This is in keeping with the analysis using the life table methods in (a).

(e) The life table conditional net survival estimates and their confidence limits are moved to different variables, depending on the level of the stratum variable. This simplifies the plot that compares the regression and life table estimates.

(f) estimates are combined into a single dataset and plotted. Note that only the non-missing rows from the *\_events\_* file (for the regression estimates) and the *conditional* file (for the life table estimates) are selected for plotting. The resulting plot is copied here.



## Q101 Kaplan-Meier (life table approach)

**Data**  Colon (a sample dataset with 35 case records)

**Procs**  lifetest

**Comments**

This exercise uses the cause of death codes in the colon dataset to compute estimates of cause-specific death. All non-cancer causes of death are considered as censored events. survival life tables are introduced, using proc *lifetest.* Patients at risk are presented both in the life table and the corresponding plot. This exercise introduces the life table approach to the estimation of survival, and presents both tabular and plotted output. The ‘*notable’* directive to *proc lifetest* is used in the final call to suppress the tabular output.

## Q102 Kaplan-Meier (life table and product-limit approach)

**Data**  melanoma (stage 1)

**Procs**  lifetest

**Comments** life table of a larger dataset. Product-limit estimates for both month and years. Plots show use of number at risk default in *lifetest*. Note use of the ‘nocensor’ option in the plot request, which turns off the default censoring indicator in the plotted survival curve

## Q103 survival and hazard functions

**Data**  melanoma

**Procs**  lifetest freq format print sgpanel

**Macros**  %mrate

**Comments** example survival and (default) smoothed hazard plots. Since the event in the survival analysis is death, the hazard rate is interpretable as a mortality rate within the cohort (i.e., using the person-time in the cohort as the rate denominator).

Macro *%mrate* can be used to compute mortality rates based on the survival data at hand. A scale factor (optional parameter ‘per’) can be specified to make the rates more readable. Confidence limits are provided. The default level of 95% confidence limits may be changed with the optional ‘level’ parameter.

Proc *sgpanel* is used to present ‘all causes’ and ‘cancer only’ plots side by side so they can be more easily compared.

## Q104 Kaplan-Meier survival and statistical tests

**Data**  melanoma

**Procs**  lifetest print

**Macros**  %mrate

**Comments** *lifetest* with strata showing statistical tests associated. *%mrate* showing mortality rate per 1,000 person time averaged over the follow-up time. Shows use of ‘per’ parameter and ‘scale’ parameter. ‘Per’ controls the rate per event time, while ‘scale’ converts between relevant scales (scale = 12 to convert survival in months to survival in years, scale = 365.24 is used to convert survival in days to survival in years)

## Q110 Mortality rates and poisson regression

**Data**  diet

**Procs**  genmod univariate sgplot print means

**Macros**  %mrate

**Comments** *%mrate* first shows that there is some relation between energy intake and mortality. The mortality rate ratio from this analysis (rate in the high energy group compared to that in the low energy group) is 7.0748 / 13.596 = .5204. That is, the high energy group experienced about ½ of the mortality rate of the low energy group.

Proc *genmod* is then used to compute fit a poisson model to the rate of mortality in the two groups. The exponentiated parameter estimate for the high energy group (compared to the reference low energy group) is .5304. This illustrates the use of poisson rate regression to estimate a rate ratio. Mortality rate ratio in this case. The idea is further explored with energy intake as a 3-level class variable, with the levels chosen. Analysis can be either with dummy variables or using the categorical variable ENG3 as a class variable.  
  
Finally, the total mortality rate is computed manually by summing deaths and person-time and compared with the results of the unstratified use of the *%mrate* macro.

## Q111 Mortality rates and poisson regression with interactions

**Data**  melanoma (restricted to stage 1)

**Procs**  genmod lifetest

**Macros**  %mrate %stset %lexis

**Comments** Kaplan-Meier survival and smoothed hazard plot (*lifetest*) stratified by period of diagnosis is used to examine the relationship to period of diagnosis. %mrate prints the overall mortality rate in the two strata.   
  
macro *%stset* is introduced to show how it creates an analytic dataset for use in other survival macros used in survival regression. This macro is unneeded for the rest of the exercise.  
  
poisson regression (*genmod*) estimates the mortality rate ratio that should replicate the ratio from %mrate.   
  
Survival is truncated at 10 years of follow-up, and the *%lexis* macro is used to compute the contribution of each subject to 1-year follow-up intervals. *%Mrate* presents the mortality rates in each interval and a plotted (*sgplot*). A smoothed version of the way cancer mortality rates change over follow-up time is plotted from lifetest.   
  
this data is further explored with a series of poisson models, showing how regression estimates of the hazard ratio (mortality risk ratio) can be adjusted for follow-up time, sex, age groups and interactions between covariates (*genmod*). Note likelihood ratio tests (type 3 analysis) for inclusion of intervals, main effects and interactions

## Q112 Poisson regression with attained age as time scale

**Data**  diet

**Procs**  genmod

**Macros** %hazard\_late %lexis %smooth

**Comments** alternate time scales in use: age at diagnosis/death; years from diagnosis to death/censoring. Age at event might be used to allow for fine control of the effect of age on survival. *Lifetest* would ordinarily be used to show the shape of the hazard function, but it cannot handle late entry (left truncation). When age is the time scale a special macro (*%hazard\_late*) uses proc *phreg* to compute the survival curve, and then another (*%smooth*) to estimate, smooth and plot the underlying hazard function.  
  
Poisson regression is then used in analyses of energy intake, accounting for job and BMI as potential confounders. Parallel analyses using the two alternate time scales

## Q120 Cox regression compared to poisson regression

**Data**  melanoma (restricted to stage 1)

**Procs**  genmod lifetest phreg sgplot

**Macros**  %lexis %rcsgen

**Comments** poisson regression can be used to estimate hazard ratios. Here *phreg* is compared with *genmod*. Poisson regression is modelled first with a parameter for each interval (120 intervals for monthly follow-up to 10 years), then with cubic splines (created by *%rcsgen*) to model the shape of the hazard function. The distinction between Cox regression and poisson regression is in the estimation of the baseline hazard function. Cox regression makes no assumptions about the shape, and results in a non-parametric estimator. *Phreg* does not output a hazard function, so it is estimated from differences of the cumulative hazard function. This is plotted (for a specified covariate pattern) with the spline fit estimated from poisson regression. The (log of the) Cox hazard function is smoothed with the *loess* procedure in *sgplot*.

These points are also covered in [Q130](#_Q130__Understanding), in more detail in Paul Dickman’s tutorial (<https://www.pauldickman.com/software/stata/compare-cox-poisson/>) and in excruciating detail in the tutorial from Bendix Carstensen (‘who needs the cox model anyway - Carstensen 2019.pdf’) in the documentation\background and theory folder. Read them for more comments on the benefits of modelling of the hazard function.

## Q121 Cox regression compared to poisson regression, interactions

**Data**  melanoma (restricted to stage 1)

**Procs**  genmod phreg lifetest

**Macros**  %lexis %smooth

**Comments** display hazard plot from *lifetest* (using *%smooth* macro, which seems to work better than the default hazard plot that *lifetest* produces). Log hazard plots should be parallel, if the proportional hazards assumption for the stratification variable is reasonable.  
  
Schoenfeld residuals are plotted from the *phreg* model. This allows a visual inspection of the proportional hazards assumption, and suggests that the assumption might not hold. If the PH assumption holds, there should be no obvious relation between the residual plot and time (or log time).   
  
To explore the possibility of non-proportional hazards, we fit a Cox model with period of diagnosis (**year8594**) as a time-varying parameter. First, using programming statements, allowing the parameter to vary after 2 years. Secondly, by splitting the data into follow-up up to 2 years, and between 2 and 10, using *%lexis*.   
  
Finally, fit the analogous poisson model with *genmod* on the split data.

## Q122 Cox regression for cause-specific mortality

**Data**  melanoma (restricted to stage 1)

**Procs**  means phreg

**Macros**

**Comments** using the cause of death codes in the melanoma data to fit different PH models by declaring some deaths to be censored.

## Q130 Understanding splines

**Data**  melanoma

**Procs**  means genmod sgplot

**Macros**  %lexis %rcsgen

**Comments** This exercise provides an introduction to the use of splines to describe the baseline hazard function. A spline in the case of survival analysis is a term for a function of a continuous variable that is used to describe the relationship between that variable and an outcome. Here, a variety of splines (linear, quadratic, cubic) are used to represent the relation between hazard (mortality) rate and time since diagnosis. The data is grouped by covariates (age, sex, period of diagnosis) but the focus is on the baseline hazard function, that is, without any covariate control.

(a) truncate the follow-up to 10 years and split the data into monthly intervals using the *%lexis* macro. This macro creates 120 records for each input record. Each individual record describes the time contribution of the subject to a single follow-up interval and includes a death/censored flag. A validity flag (\_st) can be used to exclude intervals after the subject dies. These records are created by default, but are irrelevant for the purposes of this analysis.

Proc means is used to collapse across all subjects, retaining one record for each interval and for each combination of the covariates. Compute a variable representing the value of follow-up time at the center of each interval. Some sample intervals are displayed.

(b) use poisson regression in *proc genmod* to compute a risk estimate for each interval. This results in 120 raw risk estimates, which are to be compared with spline-modelled estimates in the next steps

(c) compute the (scaled) predicted baseline hazard rate for each interval, and plot on a scatterplot against the midpoint time.  
  
(d) compute two independent linear functions (splines) of the midtime variable using a cut-point (or knot) at 2 years. Fitting these as covariates in a *genmod* model requires 4 parameters (an intercept, two linear terms and separate intercept term for the second linear spline). Plot the predictions from this model against the scatterplot of the raw interval estimates.

(e) fit a regression model without the second intercept. This results in two linear splines that are forced to join at the knot.   
  
(f) compute 2 cubic polynomials with a knot at 2 years. This requires 8 parameters. The polynomials do not join at the knot because of the separate intercept for the second polynomial.  
  
(g) by dropping the separate intercept term, the polynomials are constrained to join at the knot. This model has 7 parameters.  
  
(h) by dropping the second linear term, the curves will be continuous in the first derivative (at the knot). The resulting curve looks smoother, but is discontinuous in the 2nd derivative because of the second quadratic term still in the model.   
  
(i) dropping the second quadratic term results in a slightly smoother-looking plot, which will be continuous in the 2nd derivative at the knot.   
  
(j) use *%rcsgen* to compute restricted cubic splines with 5 knots (4 cubic splines). The knots are chosen (by default) so that there are approximately equal numbers of events (deaths) in each interval. The *fw=deaths* directive supplies the number of deaths in each interval. The first of the splines is a linear function of the time since diagnosis. The rest of the splines are cubic functions of the distance between the time since diagnosis and the corresponding knot point. They are defined over the whole range of follow-up time. The raw splines created by *%rcsgen* are highly correlated, so by default an orthogonal set of splines is returned. This behavior is turned off here, so the first spline is equal to the midtime variable. The *genmod* step fits just this first spline term.  
  
(k) add the rest of the spline terms and plot the fitted hazard function. A likelihood ratio test comparing this fit with the fit from the single spline (linear) term is highly significant. Vertical reference lines are plotted to show the location of the knots that were chosen in the previous step. Boundary knots at the observed minimum and maximum death date are chosen by default, the other knots were set at (approximately) 1.2, 2.3 and 4.4 years.  
  
(l) test different knot positions for the splines. A macro is presented for the purpose of defining different knot positions, fitting the resulting model and plotting the fitted data. The example of knots at 1, 2 and 3 years (no boundary knots) results in 2 parameters being estimated, and not a very believable fit. Adding boundary knots (0, 10) requires 2 more parameters and results in a much better fit.

## Q131 Model cause-specific mortality using flexible parametric models

**Data**  melanoma (stage 1)

**Procs**  lifetest sgplot phreg

**Macros**  %stset %stpm2 %predict %range

**Comments**

(a) Survival plot from proc lifetest  
  
(b) set up analytic dataset for use with %stpm2 and the post-fit prediction functions. Fit a null model with 1df for splines (equiv. to a Weibull model). Compare predicted survival with K-M version  
  
(c) predict hazard function from Weibull fit and compare with smoothed non-parametric hazard function. As expected, not a very good match.  
  
(d) fit a flexible parametric model with 4 splines (df = 4) and compare estimated survival and hazard functions with corresponding non-parametric versions. Match is much better now.  
  
(e) fit a Cox model with *phreg* using the period of diagnosis as the only covariate. Request HR to be reported  
  
(f) fit the corresponding flexible parametric model with 4 df for splines, requesting HR report.   
  
(g) Predict survival estimates for all subjects and merge into the original dataset. This allows us to plot the survival curves for the separate levels of the covariate. Similarly for the hazard functions.   
  
(h) Note the effect of the PH assumption, most obvious when the hazard is plotted on a log scale.  
  
(i) To look at the sensitivity of the model fit to the number of knots, we write a macro to cycle through fitting models with df = 1, 2, … 6 knots. For each model, save the fit statistics and the predicted survival and hazard functions. The estimated HR for the covariate ‘year8594’ varies from 0.89 (from the model with 1 df) to 0.78 (models with 5 or 6 df).  
  
(j) plot survival and hazard functions for comparison.

(k) this section fits 10 models, randomly selection knot locations for each one. All models have the same structure: a single binary covariate (period of diagnosis) and 5df for splines. The knots are chosen uniformly at random from centiles of the distribution of (non-censored) event times. Some combinations of knots result in models that do not converge. This can happen if the knots are too close together. Otherwise, the predicted survival and hazard functions are saved for plotting, and the estimated HRs are printed for comparison.  
  
(l) include main effects of age (group) and sex. Note that all covariates must be numeric variables.  
  
(m) display parameter estimates from both Cox and flexible parametric model. Exclude the intercept and spline terms, which are representations of the non-parametric baseline hazard in the Cox model. Parameter estimates and their analytic errors are rarely very different between these two methods of estimating HR.  
  
(n) Obtain survival predictions for specific combinations of covariates. Use the *%range* macro to define a temporary time variable for the purposes of prediction. This is known as ‘out of sample’ prediction. Also, request confidence limits for each prediction point (‘options = ci’ in the second call to *%predict*)

## Q132 Extending exercise 131 time-dependent effects in flexible parametric models

**Data**  melanoma (stage 1)

**Procs**  sgplot phreg

**Macros**  %stset %stpm2 %predict %range

**Comments**  (a) Examine Schoenfeld residuals from a Cox model including age as a class variable. These residuals should show no obvious relationship with follow-up time, if the PH assumption is reasonable. The simplest way to test is to add a model term (either using programming statements or a new covariate), and then consider the LR test of the new model against the previous model without the time-varying covariate  
  
(b) Use *%stset* to set up analytic dataset required by *%stpm2*, *%range* and *%predict*. Fit flexible parametric model with age group variables, sex and period of diagnosis, using 4 spline variables to describe the shape of the underlying hazard function. Note option (options = eform) to display Hazard ratios. Use the *%range* macro to create a temporary time variable in the analytic dataset, which will speed up prediction. Predict and plot estimated hazard functions on linear and log scales. Make a note of the log likelihood value (5017.2 with 10 parameters) associated with this model.  
  
(c) perform likelihood ratio test and print chi-square, df and p-value for the addition of age as Time-Varying Covariates (TVC) variables. *%stpm2* is called with the noprint option. You will need to examine the *\_fit\_* dataset in the work library to see the log likelihood value (4997.1 with an additional 6 parameters).  
  
(d) predict (*%predict*) the new versions of the hazard functions and plot with the corresponding functions from the previous model.  
  
(e) Hazard ratios from the Cox model and from the first flexible model will be constant over time. This is what makes the models examples of proportional hazards models. With the addition of the TVC variables, the HR for age groups is no longer constant. Predict and plot 3 hazard ratios (using the youngest age group as reference). Plot the HR for the oldest age group along with its confidence band.  
  
(f) Predict and plot differences in hazard functions; example of hazard difference with confidence band  
  
(g) Predict and plot differences in survival functions; example of survival difference with confidence band  
  
(h) fit models with 1, 2 or 3 df for each TVC and plot the corresponding HR. The model with 1df for the TVC variable accounts for only the linear component of any relationship between the covariate and follow-up time.  
  
(i) fit a model with two TVC variables (age groups and sex). The confidence limits for the HR for sex now depends somewhat on the values of the other TVC variables. Stata allows for modelling directly on the hazard scale, which apparently fixes this problem

## Q133 Extending exercise 131 alternative link functions

**Data**  melanoma (stage 1)

**Procs**  sgplot

**Macros**  %stset %stpm2 %predict %range

Comments %stpm2 allows for several different link functions through the use of the ‘scale’ parameter. The default is scale = hazards, the proportional hazards (PH) model, similar to that poisson model or the Cox model. Another option is the proportional odds (PO) model, specified by the scale parameter ‘scale = odds. The Aranda-Ordaz model family includes both these scales through the use of a model parameter, ‘scale = theta’. A fixed value of theta can be set with the ‘constheta’ parameter. ‘Constheta = 0’ forces a PH model, whereas ‘constheta = 1’ forces a PO model. Theta can take any non-negative value, but these are the only two that have an interpretation in terms of hazards or odds of death. Theta can also be an estimable parameter, by specifying the ‘scale = theta’, but no fixed theta value.   
  
(a) Fit a PH model with age groups, sex and period of diagnosis, 4 df for baseline splines and predict survival and hazard for each age group, saving estimates and fit statistics for later plotting.  
  
(b) Similarly for (a), but use the PO model  
  
(c) plot corresponding hazard and survival curves from the two models  
  
(d) print table of sit statistics  
  
(e) use %predict to compute the odds ratio (ie, using the hrnum / hrdenom parameters with a PO model) and plot  
  
(f) Compare odds ratios for different covariate patterns  
  
(g) fit the Aranda-Ordaz model with unspecified theta and display fit statistics from the differing models  
  
(h) print theta value from A-O model

## 

## Q140 Probability of death in a competing risks framework (cause-specific survival)

**Data**  colon

**Procs**  lifetest phreg sgplot sgpanel

**Macros**  %stset %stpm2 %stpm2CIF

**Comments**  A competing risks framework acknowledges the real world relationship between causes of death, where a change in the risk of one cause will induce an apparent change in the risk of an unrelated cause. This concept of competing risks is distinct from the concept of ‘relative’ or net survival, where an estimation of survival of cancer patients is considered in a hypothetical world, where cancer is the only potential cause of death. At younger ages, this is perhaps a reasonable approximation, but for older patients, the risk of non-cancer as a cause of death cannot be ignored. Net survival is appropriate for comparisons between jurisdictions, or over time, when background mortality may have changed, but is generally inappropriate for interpretation at the patient level, as it ignores non-cancer causes of death.  
  
This exercise starts by computing the product-limit (Kaplan-Meier) survival curves for cancer and non-cancer deaths, by considering one cause the ‘event’ and the other a ‘censored’ observation. That is, using the coded cause of death (cancer/non-cancer) as the only potential cause of death. These survival curves are converted to ‘failure’ curves (= 1-survival), and plotted against the competing risks interpretation from proc *lifetest*. These cumulative failure (or mortality) rates are (rather confusingly) referred to as Cumulative Incidence Functions (CIF). The effect of increasing non-cancer cause of death is seen to reduce the cancer cause of death; as non-cancer deaths become more likely, it becomes correspondingly less likely that the patient will die from cancer.  
  
Alternatives to the competing risks analysis with *liftest* are *phreg* (with the ‘eventcode’ option), which offers a non-parametric baseline estimator, and the flexible parametric analysis with *%stpm2*, followed by *%stpm2CIF*. For this latter analysis, the data is set up in a multiple-records form, with two records for each patient, one for each potential outcome. If the patient is censored at the end of follow-up, then both records record a censored event. Otherwise, one is coded censored, the other as an ‘event’, depending on which outcome happened (cancer death or non-cancer death). *%stset* is then used to create the analytic dataset from this multiple-record data. *%stpm2* offers a wide range of parameterisations of baseline hazard and choices of covariate and death cause interactions.   
  
**Fun Fact:** this colon dataset is from Finland (where Paul D. trained under Timo Hakulininen), where there was enough doubt about the accuracy of dates of diagnosis, that survival analysis usually used months as the time scale.  
  
(a) Kaplan-Meier (proc lifetest) to estimate cause-specific survival. For this, consider the alternate cause of death to be a censored event. Stratification by sex allows for independent estimation. Two runs are necessary, one for each cause of death. In keeping with the presentation of competing risks functions, a cumulative failure estimate is computed from the cumulative survival estimate.  
  
(b) *Lifetest* can also compute a non-parametric estimate of the CIF directly (using the ‘failcode’ option). Stratification by sex results in four CIF estimates.   
   
(c) Stratification is the only option for examining the effects of a covariate, when using the Kaplan-Meier method. This step computes CIF by age strata.  
  
(d) Stratification and plot by stage strata  
  
(e) Another option for a non-parametric CIF analysis is to use *phreg*. This step uses sex as a covariate (as opposed to a stratum variable). So here, the effect of sex is proportional on the ‘subhazard’ scale. For each phreg run, the ‘eventcode’ option defines the cause of death code (of the status variable) that is of interest. All other non-censoring codes are considered to be competing events. This step computes the cancer CIF, in the presence of non-cancer competing causes. Separate CIF estimates by sex can be generated in two ways: the estimated HR can be used to adjust the male CIF; a separate baseline CIF can be generated using an ancillary file to specify a covariate pattern of interest- one CIF will be generated in the baseline out file for each covariate pattern specified. These two methods will produce slightly different results.  
  
(f) Using the second method, generate the CIF for non-cancer deaths, considering cancer deaths as a competing cause. Plot the two CIF estimates from (e) with these.  
  
(g) (i) construct the dataset required by %stpm2 to model competing risks data and review the data created for a few cases. Note the censoring indicator (‘event’) as well as the indicator for cause of death (‘cause’). ‘Status’ is the variable from the original dataset, which has levels for both causes of death, as well as two censoring values. Two new variables are created, one for each cause of death.  
  
(g) (ii) use %stset to build the analytic dataset, and fit the competing risks model with %stpm2. The covariates are now sex, and the two cause indicator variables. Note that a separate set of splines is used to describe the baseline hazard function for each cause. This is accomplished by declaring the ‘cancer’ and ‘other’ variables to be time varying covariates (TVC). The default baseline splines are turned off with the ‘rcsbaseoff’ option, as they are not needed. 4 splines are used for each cause. Note also the ‘noint’ option which turns off the default intercept term, as there is now a separate intercept for each cause.  
  
(g) (iii) the impact of sex on the CIF estimates in (ii) is the same. To allow for sex to have a different effect by cause, define some new variables to include in the model. There is no longer a ‘main’ effect of sex, as there are now two parameters, one describing the effect of sex on cancer deaths, and one for the sex effect on non-cancer causes. And compare likelihood ratios to test for the importance of the effect.   
  
(h) Use *%stpm2CIF* to predict the CIF functions for cancer and non-cancer death by sex, and compare with the estimates from the earlier *phreg* (non-parametric) analysis.   
  
(i) An alternate display of CIF functions can be obtained by stacking the estimates and using the ‘band’ plot type in *sgpanel*. The total of the CIFs is identical to the ‘crude’ cumulative failure function. The CIFs are therefore a partitioning of ‘all causes’ of death into ‘cancer’ and ‘non-cancer’ (based on cause of death coding) in this competing risk framework, which is the ‘real world’ of the cancer patient.  
   
(j) To investigate if age has a different effect depending on cause of death, set up age by cause interaction variables and fit the flexible parametric model.  
  
(k) Predict and plot the age-specific CIF functions to the youngest and the oldest age groups  
  
(l) Add time-dependent effects for cancer by adding sex-cancer and age-cancer effects as TVC variables with 3 splines each (total of 12 new parameters).   
  
(m) consider if knot placement should be different for the two different death causes. In the models to this point, 4 splines were fit for each baseline function, but the knot locations are determined by the cumulative distribution of *all* death times. Here we look at the distributions separately, and fit separate models to the two death causes by partitioning the dataset into the rows that represent those causes, then saving the knot positions for each.  
  
(i) the histogram of event times seems to show different patterns of event times. This makes sense, as most cancer-related deaths occur early in the follow-up period, while non-cancer deaths, although also showing an early peak, are more uniformly distributed.  
  
(ii) Note that a temporary version of the analytic dataset must be created first, with just the cancer deaths. The full dataset is saved for later use. Fit the model to the cancer death data and copy the knot positions for both baseline and tvc variables by reading (into macro strings) from the \_model\_ database that describes the model just fit. Proc *sql* is very handy for this.  
  
Do the same for the non-cancer death records. Note that there are no non-cancer variables that are in the TVC list, so the SQL code to save the knot positions is simpler.  
  
(iii) Recover the full dataset and fit the model with cause-specific knot positions. Compute new versions of the CIF functions and combine with the earlier TVC model. Plot the youngest and oldest age groups to compare. Also, compare AIC and BIC between the two models.  
  
(n) the rest of this exercise shows how to compute CIF from a stratified Cox model, using the same ‘expanded’ data layout as in the flexible parametric version. This would allow for CIF estimates to be computed with maximum flexibility of specification of covariate by death cause interactions. While this may be illustrative, it is rather complex, and not entirely complete. Sections (o) and (p) from the corresponding Stata exercise on which this is based are ‘under development’.

## Q201 Life table estimates of relative survival

**Data**  melanoma (stage = 1)

**Procs**  sgplot

**Macros**  %rel\_surv

**Comments**  %rel\_surv is the updated version of Paul Dickman’s original sas program for computing life table estimates of relative survival. It is in use by Stats Can and PHAC in their reporting, all the survival estimates in the Canadian Cancer Statistics series. This exercise shows some of the functionality of %rel\_surv, all in the context of the cohort analytic approach: differing specification of intervals, selection of output, calculations based on results from the program. The program has been enhanced to add the Pohar-Perme net survival estimator, and the estimates of crude probability of death (competing risks estimates) based on the work of Cronin & Feuer.

a) estimates of relative survival (Ederer 2) with annual intervals:  
*intervals = 0 to 10 by 1*

b) half-year intervals  
*intervals = 0 to 10 by .5*

c) 3-month intervals for first year, followed by annual intervals (note use of %str () macro quoting function to ‘hide’ the imbedded commas  
*intervals = %str(0, .25, .5, .75, 1 to 10 by 1)*

equivalent to   
*intervals = %str(0 to 1 by .25, 2 to 10 by 1)*

d) annual intervals to 20 years  
*intervals = 0 to 20 by 1*

e) several approaches to plotting the estimates from the default file of the collapsed data (‘*grouped*’). Note that the Ederer 2 estimates are held in the variable *cr*, with 95% confidence limits *lo\_cr* and *hi\_cr*

f) plot interval-specific relative survival estimates. Note that the relative survival for intervals after 5 years starts to approach 1. That is, most of the effect of cancer mortality has diminished. A similar plot can be obtained for the net survival estimates (the Pohar-Perme interval estimates are held in the variable *ns\_w*). Notice also the variability that is a feature of this estimator.

g) report Ederer 2 and Pohar-Perme estimates using the ‘list = ‘ directive for side-by-side comparison

h) Perform ‘manual’ calculation of interval-specific rates and plot the observed (Ederer 2) excess mortality by time since diagnosis. This reiterates the point that the rate of death in the ‘hypothetical world’ of relative survival decreases after 5 years, and offers some justification for computation of ‘statistical cure’ ([Q261](#_Exercise_Q261_)) and the reporting of conditional survival from life tables ([A103](#_Exercise_A103_conditional)) or from regression modeling ([Q231](#_Exercise_Q261_), [Q231](#_Exercise_Q231_))

## Q202 Life table estimates of cause-specific survival

**Data**  melanoma (stage = 1)

**Procs**  sgplot lifetest

**Macros**  %rel\_surv

**Comments**  this program provides an example of comparing cause-specific survival (cancer-specific) with relative survival. The melanoma data is loaded, and indicator variables for cancer deaths and all deaths are coded, just for the sake of simplicity in tabulation. In the following steps, cancer or all causes is chosen by specifying what events are censored  
  
(a) use *%rel\_surv* to compute and display crude cause-specific death rates and cumulative survival probabilities  
  
(b) compute the same information with *lifetest*.  
  
(c) use %rel\_surv to compute relative (and net) survival. The only difference is in the definition (through the ‘censor’ parameter) of which events are censored. All estimates, including Ederer II and PPE are retained in a new dataset defined by using the ‘crude\_estimates’ parameter. Otherwise, the default dataset (‘grouped’) would be created, which might be overwritten by another %rel\_surv run. These estimates are combined with the cancer-specific survival estimates from (a), tabulated and plotted. Net survival and relative survival are typically larger than cancer-specific survival. Net survival tends to be somewhat more variable than E2 estimates, because of the (inverse probability of death) weighting involved in its calculation.

## Q203 Period and complete analysis I

**Data**  melanoma (stage = 1)

**Procs**

**Macros**  %rel\_surv

**Comments**  this program provides an example of using %rel\_surv to compute relative survival based on dates. When using dates for period analysis, three dates are required: origin (date of diagnosis); entry (start of period window); exit (date of death or censoring). For complete analysis, only origin and exit are required. Note that, since dates are used to specify length of survival, the ‘scale’ parameter (scale = 365.24) is needed to convert the analysis scale to years.

## Q204 Period and cohort analysis II

**Data**  melanoma (stage = 1)

**Procs**  sgplot sgpanel

**Macros**  %rel\_surv

**Comments**  Similar to exercise 203, this program computes relative survival, based on a true cohort approach (where all subjects have a potential survival of 5 years), period approach (for cases alive in a specific year) , and the actual observed survival for cases diagnosed in the specified year. It is an example of the analysis that could be done to confirm that period analysis is a reasonable predictor of actual (future) survival. Final plot of all three approaches to calculate 5-year relative survival for a range of years  
  
load the melanoma data restricted to stage 1 diagnosed before 1984.   
  
(a) traditional cohort estimates, all cases diagnosed prior to 1984. All subjects have at least 9 years of follow-up.  
  
(b) traditional cohort estimates, cases diagnosed 1977 – 1983  
  
(c) period estimates, period window 1 Jan 1983 to 12 Dec 1983  
  
(d) period estimates, period window 1 Jan 1982 to 12 Dec 1983  
  
(e) actual survival estimates of patients diagnosed in 1983  
  
(f) actual survival estimates of patients diagnosed in 1984  
  
(g) relative survival for all three methods, cases diagnosed 1981 – 1999, plot 5-year survival estimates with confidence limits on same graph for comparison

## Q210 Modelling excess mortality using Poisson regression I

**Data**  melanoma (stage = 1)

**Procs**  nlmixed genmod (with user link function)

**Macros**  %rel\_surv

**Comments**  Modeling of excess mortality (ie, regression analysis of relative survival) with poisson regression has been mostly superseded by the use of the flexible parametric survival modeling of excess mortality. This program demonstrates the methods that could be used instead of *%stpm2*. An initial run of %rel\_surv creates interval-specific estimates of excess mortality and person-time at risk. User-written link functions for genmod, or user-written likelihood functions maximized by proc nlmixed or proc nlp, can be used to estimate regression parameters.

## Q211 Modelling excess mortality using Poisson regression II

**Data**  melanoma (stage = 1)

**Procs**  iml genmod (with user link function)

**Macros**  %rel\_surv %rcsgen

**Comments**  Modeling of excess mortality (ie, regression analysis of relative survival) with poisson regression has been mostly superseded by the use of the flexible parametric survival modeling of excess mortality. This program demonstrates more specific methods of fitting excess hazard models using *%rel\_surv* (to compute the interval-specific excess hazards) and *genmod* to fit the model. Proc *iml* is used to compute interval estimates. This code is very fragile, and offered for illustrative purposes only.  
  
(a) Split the follow-up time into monthly intervals, using %rel\_surv to compute excess hazard and relative survival estimates for each interval. Results are saved in the default output file (‘grouped’) for use later. No printed output is produced.  
  
(b) Sex is recoded to be a 0/1 variable, and *%rcsgen* is used to compute 5 spline functions with pre-specified knot locations for fitting in the *genmod* model.  
  
(c) predict and plot the excess hazard rate. Genmod predicts the (log of the) fitted count. To compute a rate, divide by the person-time at risk in the interval. Note the plot on the log scale, which demonstrates   
  
(d) fit time dependent effects of the age variables by computing new interaction terms between the age and spline variables. A macro is provided to generate the 20 lines of code required.  
  
(e) compute the excess hazard ratio (and 95% CI) with youngest age group as reference. Plot estimates for specified covariate pattern (female, diagnosed in later time period).  
  
(f) plot excess hazard ratio for oldest age group with confidence limits.  
  
(g) Generate new time variable for estimation. IML is used to compute the relative survival estimates and 95% CI from the parameters of the excess hazard model last fit.

## Q230 Modelling excess mortality using flexible parametric models I

**Data**  melanoma

**Procs**  sgplot

**Macros**  %stset %stpm2 %predict %range

**Comments**  Modeling of excess mortality (ie, regression analysis of relative survival) with %stpm2. Examples of use of %predict for hazard, survival, hazard ratio, hazard difference, survival difference. Evaluation of range of choices of numbers of spline functions for baseline (excess) hazard.  
  
(a) load data and append mortality rates for each subject. Rate chosen is the general population mortality rate matched to the patient’s age at time of death/censoring. Use *%stset* to build the analytic dataset and *%stpm2* to fit base model (no covariates).  
  
(b) *%predict* and plot hazard and survival functions from (a) model. Note that *%predict* saves all predicted estimates as new variables in the \_events\_ dataset. Since this is an excess hazard model, the estimates are predicted (fitted) *excess* hazard and *relative* survival.  
  
(c) fit alternative models using 2, 4, or 6 spline curves. Plot excess hazard and relative survival functions and compare fit statistics. 2df is not enough, 4 is better and 6 is probably over-fitted. Note that a likelihood ratio test is not appropriate, because the models are not nested. AIC and BIC are both penalized likelihoods and are used to evaluate these models.  
  
(d) fit excess hazard model with age group, sex and period of diagnosis. This is a proportional excess hazard model. %stpm2 reports the hazard ratio because of the ‘eform’ option. %Predict survival and hazard functions, and scale the hazard function with the ‘per’ option to be able to interpret excess hazard per 1,000 person years.   
  
(e) plot survival estimates; plot hazard estimates with a log scale, showing effect of proportional modeling.  
  
(f) add time-dependent effects for age group (2 splines for each age group). A likelihood ratio test for the inclusion of these terms is appropriate, since the models are nested. 6 new parameters have been added, and the difference in -2log likelihood is 12 (16971 – 16959). This is a chi-square with 6df which has an associated p-value of .06. %Predict and plot the new versions of the survival and hazard functions.  
  
(g) use the *%range* command to create a temporary time variable reduce the number of data points to estimate (401 instead of 7,700). This strategy speeds up prediction steps and is often a good idea. *%predict* and plot the excess hazard ratios for each age group with the youngest as reference.  
  
(h), (i) %predict and plot differences in relative survival and differences in excess hazard with 95% confidence bands

## Q231 Modelling excess mortality using flexible parametric models II

## 

**Data**  colon

**Procs**  sgplot loess tabulate

**Macros**  %stset %stpm2 %predict %range, also some purpose-built macros for repetitive tasks in the program

**Comments**  Fit and evaluate models of relative survival. This is a long program, taking more than 14minutes on a lower-powered Windows 7 machine with 8Gb memory and an AMD A8-3870 processor running at 3Ghz. The code for timing the whole program is commented out in the exercise.  
  
(a) load data and fit base model with no covariates. All causes of death are used here, since we are interested in relative survival.   
  
(b) *%predict* the martingale residuals for the model, and plot against patient age with *loess*. Martingale residuals are the difference between the observed and the fitted deaths. The plot against a covariate helps evaluate if the functional form of the covariate is being modelled well, and the loess fit provides an overview of that relation.  
  
(c) add age as a linear covariate  
  
(d) estimate the excess hazard ratio for each patient at their stated age. The ‘at = age:.’ Directive requests an estimate at the stated age without change. The reference age (50) is used in the ‘hrdenom’ parameter. Plot the effect of age on excess mortality rate ratio. Since the estimates are saved in the analytic dataset (\_events\_), they must be extracted and sorted by age for plotting.   
  
(e) *%predict* the martingale residuals again, and plot against age at diagnosis with *loess*. There is still a considerable relationship between age and the residuals, which means that using age as a linear term is not sufficient.   
  
(f) generate splines for age and fit new model with age splines instead of age as a linear term. Note that we save the knot points for the age splines, and the orthogonalization matrix, which will be needed later on to do predictions by age in the new model. The loess fit of the martingale residuals is better now, with 4 df for age, rather than one. The likelihood ratio test of model fit of the spline version of the model versus the linear version of age is highly significant. chi square of 103 (=36220 – 36107) with 4 df.   
  
(g) If there is a disadvantage to using splines to describe the relation between survival and age, it is this: it becomes difficult to use *%predict* for specific ages, as the program needs the spline values – not the age value itself – to be able to predict. Here is one way to accomplish this. *%rcsgen* is used to compute the spline values from the knots and orthogonalization matrix saved earlier. The spline values are loaded into macro strings with *sql* code, and inserted into the *%predict* calls. This is a fairly general method, and could be used as a template elsewhere if  
 of range of choices of numbers of spline functions for baseline (excess) hazard. But it is a little tricky…  
  
(h) here is another way to evaluate a measure over a range of ages. We estimate the one-year relative survival for every patient in the analytic dataset by creating a new temporary time variable that has the value of interest (1 year) for each row. The *%predict* call uses default covariate values for the prediction, that is, each patient row is evaluated without changing the value of the covariate (there is no ‘at = ‘parameter in the call)  
  
(i) similarly, for a plot of 5-year relative survival by age at diagnosis  
  
(j) survival at 5 years, conditional on survival to 1 year, by age. This makes use of the *tcond* parameter, which supplies the time already survived in the scale currently in use. For relative survival regression, this will typically be years. With the confidence intervals requested (*options = ci*), this step takes a minute or so to run.  
  
(k) Estimate the excess hazard ratio by age, with 50 as the reference age. The macro that performs the loop through the ages of interest must first get the spline values for the reference age, and store them in macro strings. Then, within the age loop, the spline values are computed for each individual age for the call to *%predict*. Note that at age 50 the HR is 1, since the numerator and denominator ages are identical. This means also that the SE of that value is zero, and also that for ages near 50, the SE will also be small.  
  
(l) perform the same HR by age analysis, with 3 different model definitions, using 3, 4, or 5 spines for the description of the age relation. This needs the previous macro to be embedded within a loop that cycles through the three models to be used. Finally, compare the plots visually, and the fit statistics in a table. This particular step takes a while to run.

Q232 Modelling excess mortality using flexible parametric models III

**Data**  colon

**Procs**  sgplot sgpanel sql

**Macros**  %stset %stpm2 %predict %range. some purpose-built macros for repetitive tasks in the program.

**Comments**  Fit and evaluate models of relative survival. This takes about 5 minutes to run. The code for timing the whole program is commented out in the exercise. Note the age restriction. It might be a policy decision to exclude the very oldest patients.   
  
(a) load data and append population mortality rates; build analytic dataset.   
  
(b) define 4 splines to use in fits with *%rcsgen*. Fit base model with age splines and note the -2 log likelihood value   
  
(c) add time-dependent factors for the age splines, amounting to 8 new parameters. Since the previous model is nested in this one, the LR test is appropriate.  
  
(d) set a temporary time variable with *%range* and use macro to *%predict* excess hazard and survival functions at selected ages.  
  
(e) Use the same methods as in 131 to *%predict* 1 year survival as a function of age. The method used here is to predict 1-year survival for every patient in the dataset (15,000 rows). Confidence intervals require intensive numeric methods because of the need for first derivatives of the prediction function, so an alternate method would be to restrict the \_events\_ dataset to rows representing the distinct covariate combinations (76, in this example). The estimation for all 15,000 rows only takes a minute or so.  
  
(f) 5-year relative survival by age at diagnosis. When no ‘at=’ parameter is present in the call to *%predict*, covariates are not changed from their observed values.  
  
(g) 5-year relative survival, conditional on survival to 1 year already lived. Note use of the ‘tcond = 1’ parameter in the call to *%predict* to specify that 5-year survival is conditioned on having already survived 1 year. If confidence intervals were not required, (that is, do not specify ‘options = ci’) processing would be much faster. Or the conditional estimates could be obtained by dividing the 5-year survival relative survival estimate by the 1-year relative survival estimate.  
  
(h) Use the *%range* command to create a temporary time variable of points at which to estimate the excess hazard ratio. Loop through 4 selected ages and plot  
  
(i) predict hazard differences and survival differences at 4 selected ages  
  
(j) evaluate number of df for TVC models. %predict HR for age 70 vs age 50. Print fit statistics

## Q240 Age standardization life table methods I

**Data**  melanoma (stage 1)

**Procs**  sgplot tabulate

**Macros**  %rel\_surv .

**Comments**  This exercise uses the %rel\_surv macro to compute age-standardised relative survival estimates. The final plot shows Ederer II and the Pohar-Perme (net) age-standardised survival estimates on the same plot.  
  
(a) load the data and add an age-specific weight variable for use in a later step. Use %rel\_surv with survival time in months, and then again with dates. Resulting estimates should be identical.   
  
(b) run *%rel\_surv* with age group as a stratifier. Use *freq* to get the age-specific percentages from the melanoma data to use as weights. Merge the percentages from *freq* and scale to a proportion to the age-specific estimates for 10-year survival from the *%rel\_surv* run (using the default ‘grouped’ file), and use them to weight the estimates. Finally, use proc tabulate to sum across age groups to get the (internally) age-standardised estimate. Is the cumulative relative survival in the ‘All’ row meaningful?  
  
(c) %rel\_surv run with the user-supplied weights from (a). This method has the benefit of producing a confidence interval, whereas the manual method (b) can only replicate the point estimate (if the weights are the same).  
  
(d) %rel\_surv computes and saves both Pohar and Ederer II as crude estimates in the ‘crude\_estimates’ file (default name is ‘grouped’). These are plotted for comparison.

## Q241 Age standardization life table methods II

**Data**  melanoma

**Procs**  sgplot tabulate

**Macros**  %rel\_surv .

**Comments**  This exercise uses the %rel\_surv macro to compute age-standardised relative survival estimates. The final plots show Ederer II and the Pohar-Perme (net) age-standardised survival estimates on the same plot. A second plot shows the SE estimates side by side.  
  
(a) %rel\_surv gives estimates stratified by time period of diagnosis.  
  
(b) age and time period stratified analysis using %rel\_surv. Estimates for 10-year survival are age-standardised to the internal age proportions and printed.  
  
(c) Standardised estimates by user-supplied weights. Print and plot E2 and PPE for comparison. Note plot of corresponding SE estimates for the annual intervals. The PPE estimates are offset by a small amount for clarity. The PPE SE estimates are slightly wider, especially at 10 years for the later interval, where the data is thinnest. This is a known ‘feature’ of the Pohar-Perme estimator.

## Q242 Age standardization using flexible parametric models.

**Data**  melanoma (stage 1)

**Procs**  sgplot tabulate

**Macros**  %stpm2 %predict .

**Comments**  This shows several methods for age-standardising relative survival estimated from a flexible parametric survival model  
  
(a) load data, append population death rates and fit a base model (no covariates) with *%stpm2*. Use *%predict* to get ‘all ages’ relative survival. Retain estimates to be combined with age-specific results later.  
(b) fit *%stpm2* model with age categorical variables, *%predict* survival by age group and plot with the ‘all ages’ result from (a). The ‘all ages’ estimate is very similar to the age group ’45 – 59’.  
  
(c) use proc *freq* to get the age distribution for entire cohort, and use the observed proportions to create a weighted average survival estimate. This is an internally standardized estimate.  
  
(d) display the estimated 10-year age standardized estimate.  
  
(e) Now use the ‘meansurv’ option in %predict, which computes the population average survival curve. Note that this differs from the predicted survival curve at the mean of all the covariates in the model. A predicted survival curve is obtained for each subject and all the survival curves are averaged. The process can be computationally intensive, hence the sue of the temporary time variable  
  
(f) report confidence intervals computed in (e) for an example time.  
  
(g) fit a proportional excess hazard model including time period. We use *%predict* and proc *means* to examine the difference in average survival. Is there variation in survival by time period? If so, age standardization could be used to validly compare survival across the time periods  
  
(h) examine the change in the age distribution by time period  
  
(i) *%predict* using the ‘meansurv’ and ‘at=’ options to obtain population averaged survival curves. By restricting the dataset to those in the earlier time period (with the ‘ifp = ‘ option), the relative survival estimates are standardized to the age distribution of the cases diagnosed in 1974-85.   
  
(j) prediction using the age distribution in the later time period using the ‘meansurv’ option, estimated for the value of year8594 = 1 with the ‘at = ‘ specification. Plot all these survival functions and print the estimate at 10 years.

.

## Q243 Age standardization life table methods III (ICCS weights)

**Data**  melanoma (stage 1)

**Procs**

**Macros**  %rel\_surv .

**Comments**  This shows several methods for age-standardising relative survival estimated from life tables with the %rel\_surv macro. The macro provides for both user-supplied and built-in weights. The built-in weights reference two weight files, one based on the ICSS weighting system, and one with weights derived from Canadian sources.  
  
(a) load the melanoma data, create internal weights from observed age distribution. Apply these weights to the case data and run *%rel\_surv* with the internal weight options (‘weight\_var’ and ‘standstrata’).  
  
(b) append the ICSS(2) weights to the case data and run the same age-standardisation run (with the new weight variables), then use the internal library to assign the ICSS (2) weights (using the ‘stnd = icss(2)’ directive).   
  
(c) compare the weight sets assigned and the corresponding age standardized results at 5 years  
  
(d) *%rel\_surv* run requesting standardization with the ICSS(1) weight set and print a report of the different results at 10 years.

## Q250 Probability of death in a competing risks framework (life table methods)

**Data**  melanoma (stage 1)

**Procs**  sgpanel

**Macros**  %rel\_surv .

**Comments**  Use the %rel\_surv macro to compute the cumulative incidence functions using the Cronin & Feuer methods. Compute net failure (1-net survival) and overall mortality (1- observed survival) for comparison  
  
(a) load data and run the *%relsurv* step, stratified by age group and sex. Request the interval-specific report of the Pohar estimates of observed, expected and net survival, along with the crude probabilities of death due to cancer, and to other causes. These fields are always present in the saved output file (‘grouped’) but are requested with the ‘list = ‘ option. The ‘crude=’option in this case selects only the intervals ending in integers for each stratum. The ‘grouped’ file has results for all interval and strata combinations.  
  
(b) compare the death probability due to cancer with the all causes probability, for the youngest age group. These are very similar.  
  
(c) crude probabilities and all-cause mortality  
  
(d) proportions of deaths due to cancer, due to other causes  
  
(e) plot overall, net and crude probability of cancer

## Q251 Probability of death in a competing risks framework (regression methods)

**Data**  melanoma

**Procs**  sgplot

**Macros**  %stpm2 %predict %stpm2CM %range .

**Comments**  Use the relative survival regression framework provided by *%stpm2* to compute crude probabilities of death due to cancer / other causes, using the principles of competing risks.   
  
(a) load data, assign age group binary variables and assign population mortality rates from the popmort file  
  
(b) fit an excess hazard model with age groups as TVC using *%stpm2*. Note degrees of freedom for baseline excess hazard and for tvc variables. 18 parameters are needed.  
  
(c) Use the *%predict* macro for failure functions for each age group and plot.   
  
(d) The prediction function %stpm2CM is designed to work with an excess hazard model to compute the estimate of crude mortality (due to cancer / other causes). This does not require cause of death to be coded (as is needed by *%stpm2CIF* in exercise 140). Expected mortality (ie, the hazard function for ‘all-cause mortality’) is computed by an implementation of the Ederer II method within the *%stpm2CM* program. The ‘at = ‘ directive assigns the covariate value to use for the prediction of expected hazard. For the all causes (expected mortality) computation, it is also necessary to provide an age, sex and year of diagnosis.  
  
(e) In the same way as was used in exercise 140, create the stacked graphs showing the cumulative incidence functions (crude probabilities of death).  
  
(f) For the adventurous, fit an *%stpm2* model with splines for age instead of age categories. Note the improvement in fit by comparing AIC or BIC with the previous model. This also allows for finer control of the age effects. For the age-specific calls to *%stpm2CM*, the spline values that correspond to the age selected are computed by *%rcsgen*, then captured as macro strings by *sql* code.

## Q261 Cure models

**Data**  melanoma (stage 1)

**Procs**  sgplot sql

**Macros**  %stpm2 %predict .

**Comments**  The cure model implemented with *%stpm2* is a non-mixture cure model, with a special choice of splines designed to estimate an excess hazard function that approaches 0 after enough follow-up time. At least 10 years of follow-up is recommended. Since the excess hazard is 0, the remaining patients are experiencing the same mortality rate as the general population, and can be considered to be ‘statistically cured’. As is true of all cure models, a positive cure proportion at the population level will always result, even if there is no reasonable chance of a clinical cure. Predictions that are specific to this model are the cure proportion (see the ‘cure’ directive in *%predict*) and the median survival of the ‘uncured’ (use the ‘cent = 50’ directive to estimate the 50th percentile, and the ‘uncured = 1’ directive).  
  
Note on processing time: the prediction of centiles of the survival function is carried out on each row of the analytic dataset, but is only required for distinct covariate patterns. Programming statements could be included to change from a full dataset (for fitting the model) to a dataset restricted to distinct covariate patterns (for example, only 2 rows if the only covariate is 0/1). This would speed up processing considerably but makes the program harder to read. Calculations for confidence intervals (in particular on the prediction of medians) are especially time-consuming.  
  
(a) load the data, append population death rates and fit an excess hazard model. The ‘option = cure’ directive requests the spline functions that are required by the cure model. The default scale of log cumulative hazard (‘scale = hazard’) is the only model that is allowed with the ‘cure’ option. All other model specification terms are allowed. Note that the final spline parameter is held at 0 in cure models, both in baseline and TVC splines as in step (c). This will be apparent when reviewing parameter estimates.  
  
(b) estimate cure proportion and median survival in the ‘uncured’.   
  
(c) expand the model by allowing for non-proportional effects of the covariate, and estimate the cure proportion and median survival in the ‘uncured’ from this new model.  
  
(d) estimate the survival function for all patients and the survival function for the uncured population. Since there is no ‘at=’ directive in the *%predict* call, estimation takes place for each subject with that subject’s covariate and time variables unchanged. Note the use of macro strings (and special formatting in the *sgplot* routines) to draw reference lines and label them. This amount of detail is probably OTT (over the top) for most applications of the method.

## Q282 Avoidable deaths

**Data**  melanoma (diagnosed 1985-94)

**Procs**

**Macros**  %rel\_surv .

**Comments**  An example of how to compute deaths in one stratum based on death rates in another. This example uses interval-specific death rates from %rel\_surv for females to compute the number of deaths that would be expected in males, if they experienced the female mortality rate in that interval. The difference between observed and this expectation can be considered to be ‘avoidable’ deaths. The method is general enough, and could be applied to some other stratification, such as stage at diagnosis.

## Q284 Residual expectation of life and years of life lost

**Data**  melanoma

**Procs**  sgplot

**Macros**  %stpm2 %stset %predict .

**Comments** To compute total years lived by all members of a cohort of cancer patients, it would be necessary to follow them until all subjects had died. However, for diseases where the relative survival appears to level off after a number of years, the excess hazard model fit by *%stpm2* can be used as a basis for extrapolation of the observed survival curve to a point in the future where the entire cohort is deceased. The cohort should have at least 10 years of potential follow-up (as in exercise 261 – cure models) so that a clear leveling off of the relative survival curve has occurred. This exercise shows how to use the *%predict* macro (after fitting an excess hazard model) to compute the expectation of life for each patient in the cohort.   
  
(a) load data and append the population death rates to fit an excess hazard model  
  
(b) define spline variables for age (4 df) and year of diagnosis (4 df), fit excess hazard model with *%stpm2* with age, sex, year of diagnosis. Include TVC terms for age and year of diagnosis. This model has 42 parameters. For a purpose like the life lost estimation, where the shape of the hazard function is of most importance (as opposed to adequate estimation of hazard ratios), we want to err on the side of over-fitting.   
  
(c) predict expectation of future life years with *%predict*, using the current popmort file. Note that the latest year in this popmort file is 2000, but we need to project much further (80 years, because of the ‘tinf = 80’ directive) after diagnosis. By specifying ‘maxyear = 2000’ all future years will make use of the population mortality rates from this final year. Estimation of the life expectancy involves integration of the survival curve. This is carried out by a numerical technique known as Gaussian quadrature. The specification of ‘nodes = 40’ is an instruction to the quadrature algorithm. Fewer nodes might give slightly less precise estimates. Note that, unlike the other prediction functions, the ‘lifelost’ estimate variables are merged back on to the analytic dataset (\_events\_). So reporting is based on that file.  
  
(d) select estimates for males (sex = 1, since the variable sex is on the analytic dataset, even though it was not specified in the model) and year of diagnosis = 1994 (the most recent data year). Plot the years of life lost against age.  
  
(e) print population expectation and years of life lost for selected ages. Note that the ‘nodupkeys’ directive in proc *sort* is used to select distinct individuals from the analytic dataset with specific covariate patterns (age restriction, year of diagnosis = 1994). This dataset is used for the report.  
  
(f) For the ‘lifelost’ directive of %predict, there is no option to specify covariate patterns of interest with an ‘at=’ parameter. To compare the years of life lost due to the male survival deficit, we force all subjects to be female, and make a second call to %predict, naming a new variable to hold the estimates. Now, each male in the cohort will have his life expectation increased because of the lower mortality rates for females. Correspondingly, his years of life lost will be decreased.  
  
(g) the difference between the true years of life lost for males, and the estimate for them as if they were females, is due to the male survival deficit. At this point, the total years of life lost (due to the male/female survival difference) is summed over the entire cohort. The specific covariate patterns of interest are selected for reporting with proc *sort*, so we can make statements about individuals.