Relative Survival Analysis using SAS

Sabuj Sarker
Department of Epidemiology and Performance Measurement
Saskatchewan Cancer Agency

October 18, 2012

The Survival Function S(t) and Hazard Function h(t)

- In survival analysis, the outcome is either the survival proportion or the event rate (the proportion who do not experience the event)

$$h(t) = f'(x) / f(x) = -\frac{d}{dt} \ln S(t) = \frac{f(t)}{S(t)}$$

- Which indicates that the event rate is proportional to the rate at which the survival function decreases
- Strictly, the hazard is the rate of change of ln *S(t)* but we can think of it as being proportional to the rate of change of *S(t)*

Prognosis of Cancer Patients

- Total mortality (among the patients)
- Net mortality (mortality associated with a diagnosis of cancer)
- Cause-specific mortality (an estimate of net mortality under certain assumptions)
- When estimating cause-specific mortality, only those deaths which can be attributed to the cancer in question are considered to be events

Cause-specific mortality

- Cause-specific mortality = number of death due to cancer/person-time at risk
- The survival times of patients who die of causes other than cancer are censored
- Need reliably coded information on cause of death
- Even when cause of death information is available to the cancer registry via death certificates, it is often vague and difficult to determine whether or not cancer is the primary cause of death

Cont..

- How do we classify, for example, death due to treatment complications?
- Consider a patient treated with radiation therapy and chemotherapy, who dies of cardiovascular disease. Do we classify this death as “due entirely to cancer” or “due entirely to other causes”?
- Some of the researcher developed an improved cause-of-death indicator

Relative Survival

- Instead of cause-specific mortality, we can estimate excess mortality: the difference between observed (all-cause) and expected mortality
  - Excess mortality = Observed mortality – Expected mortality

Cont..

- Relative survival is the survival analog of excess mortality – the relative survival ratio is defined as the observed survival in the patient group divided by the expected survival of a comparable group from the general population.
  - Relative survival ratio = observed survival proportion/expected survival proportion

Cont…

- It is usual to estimate the expected survival proportion from nationwide (or state/province wide) population life tables stratified by age, sex, calendar time, and where applicable, race
- Several different methods exist for estimating expected survival, the most commonly used: Ederer I, Ederer II, and Hakulinen

Advantages

- Information on cause of death is not required
- Obtain a measure of the excess mortality experienced by patients diagnosed with cancer, irrespective of whether the excess mortality is directly or indirectly attributable to the cancer
- Deaths due to treatment complications or suicide are examples of deaths which may be considered indirectly attributable to cancer

Estimating Relative Survival

- The relative survival ratio (RSR) is estimating using life table methods
- The cumulative RSR is estimate at discrete points in the follow-up by taking the product of interval-specific estimates over sub-intervals of the follow-up
- Sample data set containing information on colon carcinoma diagnosed in Finland

Cont..

- In general, two data files are required to estimate relative survival;
  - A file contain individual level data on the patient (colon)
  - A file contain expected probabilities of death for a comparable general population (popmort)
- Running the SAS code in survival.sas will produce the life table estimates of relative survival stratified by sex, age, and calendar period of diagnosis

Cont..

- In addition, two output data set are created;
  - One contain grouped data
  - One contain individual patient data
- These two file used for input data sets for modeling
- The SAS code in models.sas estimates a relative survival regression model using several different approaches (Described in Dickman et al.)

Overview in SAS

- Split the observation time
- To split the data, use the SAS macro 'lexis' (written by Bendix Carstensen)
- The general approach is as follows:
  - Use the lexis macro to split the observation time
  - Ensure each record has the correct value for attained age and attained calendar period
  - Merge from an external file (the so-called 'popmort' file), expected probabilities of surviving the interval

The general approach (Cont…)

- For each observation, create indicator variables for death and censoring
- Create life tables for each desired combination of covariates by collapsing over relevant records
- For each life table interval, calculate interval-specific and cumulative observed, expected, and relative survival with corresponding standard errors and confidence intervals
- Print the life table (using Proc Print)

Models for Relative Survival

- Models for relative survival can be estimated based on either individual (subject-specific) or grouped/collapsed (i.e. life table) data
- A copy of data is saved before collapsing (the default data set name is individ) and after collapsing (the default data set name is grouped)

The patient data file (colon)

- Individual-level data for 15, 564 patients diagnose with colon carcinoma in Finland 1975-1994 with follow-up to the end of 1995
- A file containing information on individuals diagnosed with cancer is required and must contain, at a minimum, the following information:
  - Survival time (time at risk)
  - Indicator for vital status (dead/alive)
  - Variables upon which expected survival depends – typically age, sex, and period

Alphabetic List of Variables and Attributes

#	Variable	Type	Len	Format	Label
2	AGE	Num	4		Age at diagnosis
11	DX	Num	8	DATE.	Date of diagnosis
12	EXIT	Num	8	DATE.	Date of exit
4	MMDX	Num	4		Month of diagnosis
1	SEX	Num	4	SEX.	Sex
3	STAGE	Num	4	STAGE.	Clinical stage at diagnosis
8	STATUS	Num	4	STATUS.	Vital status at last date of contact
9	SUBSITE	Num	4	COLONSUB.	Anatomical subsite of tumour
6	SURV_MM	Num	4		Survival time in completed months
7	SURV_YY	Num	4		Survival time in completed years
10	YEAR8594	Num	4		Indicator for year of dx 1985-94
5	YYDX	Num	4		Year of diagnosis

The population mortality file (popmort)

- Contains expected survival probabilities (PROB) for the Finnish general population stratified by age, calendar year, and sex for the years 1951 to 2000
- Calculated by the central statistical office (Statistics Finland)

Popmort Data

Obs	SEX	_YEAR	_AGE	PROB
1	1	1951	0	0.96429
2	1	1951	1	0.99639
3	1	1951	2	0.99783
4	1	1951	3	0.99842
5	1	1951	4	0.99882
6	1	1951	5	0.99893
7	1	1951	6	0.99913
8	1	1951	7	0.99905
9	1	1951	8	0.99920
10	1	1951	9	0.99931

SAS Code

- The approach is implemented as SAS code, rather than as a SAS macro, in order to make it more transparent and easier to customize
- The main parameters are, however, defined as macro variables at the top of the file: Two input data files and two output data files –

SAS Code: Cont…

- title; footnote;
- title1 'Colon carcinoma diagnosed in Finland 1975-1994 (follow-up to 1995)';
- libname colon 'H:\~\rsmodel';
- options fmtsearch=(colon work library) orientation=landscape pageno=1;
-

SAS Code: Cont…

- /***
- Define the input and output files.
- ***/
- /* Population mortality file */
- %let popmort=colon.popmort ;
-
- /* Patient data file */
- %let patdata=colon.colon ;
-
- /* Output data file containing individual records */
- %let individ=colon.individ ;
-
- /* Output data file containing collapsed data */
- %let grouped=colon.grouped ;
-

SAS Code: Cont…

- The macro variable “vars” stores the variables over which the life tables are stratified

```
%let vars = sex yydx age;
%let formats = sex sex. age age. yydx yydx. ;
```

- The lifetable may be estimated for each combination of sex, yydx (year of diagnosis), and age.
- If, for example, the variable age contains age at diagnosis in years then categories can be constructed using a format.
- Formats can be used to group metric variables into categories.

SAS Code: Cont…

- The next step is a data step where housekeeping on the patient data file is performed

```
data &individid;
   length id 5;
   set &patdata;
```

- Restrict to localized stage: local, distant, regional, unknown:

```
if stage=1;
```

- Create a unique ID for each individual

```
id+1;
```

SAS Code: Cont…

- The variable SURV_MM contains survival time in completed months.
- We will add 0.5 to all survival times, both to avoid problems with individuals with time=0 (who are theoretically never at risk and may be excluded from some analyses) and because this provides a more accurate estimate of person-time at risk (for Poisson regression analyses).

```
surv_mm = surv_mm + 0.5;
```

SAS Code: Cont..

- The lexis macro requires a variable containing the time at entry-

```
entry=0;
```

- Create an indicator variable for death due to any cause (Dead: Cancer, other) -

```
    if status in (1,2) then d=1;
    else d=0;
    drop stage subsite status;
    label id='Unique subject ID';
run;
```

SAS Code: Cont ..

Portion of the outputs:

id	SEX	AGE	YYDX	SURV_MM	d
1	Female	78	78	82.5	1
2	Male	80	80	8.5	1
3	Female	75	75	23.5	1
4	Male	77	77	85.5	1
5	Female	76	76	32.5	1

SAS Code: Cont…

- Splitting the time at risk using the lexis macro
  - Use the lexis macro to split the observation time for each individual into multiple observations, one for each band to follow-up (1-year interval, maximum 10-year for this example).

```
%include 'H:\~\lexis.sas';
```

- Split the data to obtain one observation for each life table interval for each individual. The scale must be transformed to years.

```
%lexis (data=&individ., out=&individ., breaks = %str( 0 to 10 by 1 ),
origin = 0, entry = entry, exit = surv_mm, fail = d, scale = 12, right =
right,  risk = y, lrisk = ln_y, lint = length, cint = w, nint = fu ) ;
```

SAS Code: Cont..

- In the previous code, instead of using the information on survival time in the variable surv_mm we may, alternatively, specify the date of diagnosis and exit -

```
%lexis (data=&individ., out=&individ., breaks = %str( 0 to 10 by 1 ),
    origin = dx, entry = dx, exit = exit, fail = d, scale = 365.25,
    right =    right, risk = y, lrisk = ln_y, lint = length, cint = w, nint = fu );
```

SAS Code: Cont..

- Create variables for attained age and calendar year which are 'updated' for each observation for a single individual. These are the variables by which we will merge in the expected probabilities of death, so they must have the same names and same format as the variables indexing the POPMORT file (sex, _age, _year in this example).

```
data &individ;
    set &individ;
        _age=floor(age+left);
        _year=floor(1900+yydx+left);
```

- A variable to label the life table output

```
        range=put(left,4.1) || ' - ' ||left(put(right,4.1));
        drop entry left right;
run;
```

SAD Code: Cont..

The output:

id	SEX	AGE	_age	YYDX	_year	d	w	fu	y	length
1	Female	78	78	78	1978	0	0	1	1	1
1	Female	78	79	78	1979	0	0	2	1	1
1	Female	78	80	78	1980	0	0	3	1	1
1	Female	78	81	78	1981	0	0	4	1	1
1	Female	78	82	78	1982	0	0	5	1	1

SAS Code: Cont….

- Now merge in the expected probabilities of death.

```
 proc sort data=&individ;
      by sex _year _age;
 run;
proc sort data=&popmort;
     by sex _year _age;
run;
```

SAS Code: Cont…

```
data &individid;
        length d w fu 4 y ln_y length 5;
        merge &individid(in=a) &popmort(in=b);
        by sex _year _age;
        if a;
```

- /* Need to adjust for interval lengths other than 1 year */

```
        p_star=prob**length;
```

- /* Expected number of deaths */

```
        d_star=-log(p_star)*(y/length);
        keep &vars fu range length d w p_star y ln_y d_star;
```

SAS Code: Cont…

```
label
        d_star='Expected number of deaths'
        d='Indicator for death during interval'
        w='Indicator for censored during interval'
        y='Person-time (years) at risk during the interval'
        length='Interval length (potential not actual)'
        ln_y='ln(person-time at risk)'
        p_star='Expected survival probability'
        _age='Attained age'
        _year='Attained calendar year'
        range='Life table interval'
        fu='Follow-up interval'
        sex='Sex';
run;
```

SAS Code: Cont…

- After merging in the expected probabilities of surviving (p_star), we have the following:

id	SEX	AGE	_age	YYDX	_year	d	w	fu	y	length	p_star
1331	Male	22	22	75	1975	0	0	1	1	1	0.99836
1318	Male	27	27	75	1975	0	0	1	1	1	0.99830
1321	Male	27	27	75	1975	0	0	1	1	1	0.99830
1277	Male	36	36	75	1975	0	0	1	1	1	0.99700
1278	Male	36	36	75	1975	0	0	1	1	1	0.99700

SAS Code: Cont…

- **Collapse the data to produce the life table**:

```
proc summary data=&individ nway;
        var d w p_star y d_star;
        id range length;
        class &vars fu;
/* Follow-up must be the last variable in this list */
        output out=&grouped(drop=_type_
        rename=(_freq_=l))
        sum(d w y d_star)=d w y d_star
        mean(p_star)=p_star;
        format &formats ;
run;
```

SAS Code: Cont…

Output from the group data:

SEX	YYDX	AGE	fu	range	length	l	d	w	y	d_star	p_star
M	1975-84	0-44	1	0.0 -1.0	1	75	4	0	73.42	0.22	0.996
M	1975-84	0-44	2	1.0 -2.0	1	71	8	0	66.33	0.21	0.996
M	1975-84	0-44	3	2.0 -3.0	1	63	1	1	62.50	0.22	0.996
M	1975-84	0-44	4	3.0 -4.0	1	61	3	0	59.46	0.22	0.996
M	1975-84	0-44	5	4.0 -5.0	1	58	3	0	56.54	0.23	0.996

l: # alive at the start of the interval

d: # of deaths during the interval

w: # of patients whose survival time was censored during the interval

p_star: Expected survival proportion for the interval

d_star: Expected number of deaths

SAS Code: Cont…

- Calculate the interval-specific observed and expected relative survival and then estimate cumulative survival:

```
data &grouped;
        retain cp cp_star cr 1;
        set &grouped;
        if fu=1 then do;
        cp=1; cp_star=1; cr=1; se_temp=0;
        end;
        l_prime=l-w/2;
        ns=l_prime-d;
```

SAS Code: Cont…

- /* Two alternative approaches to estimating interval-specific survival */
- /* Must use the hazard approach for period analysis */

```
p=exp(-(d/y)*length); /* transforming the hazard */
p=1-d/l_prime; /* actuarial approach */
r=p/p_star;
cp=cp*p;
cp_star=cp_star*p_star;
cr=cp/cp_star;
ln_y_group=log(l_prime-d/2);
ln_y=log(y);
```

SAS Code: Cont….

```
d_star_group=l_prime*(1-p_star);
excess=(d-d_star)/y;
se_p=sqrt(p*(1-p)/l_prime);
se_r=se_p/p_star;
se_temp+d/(l_prime*(l_prime-d)); /* Component of the SE of the cumulative survival */
se_cp=cp*sqrt(se_temp);
se_cr=se_cp/cp_star;
```

SAS Code: Cont….

- /* Calculate confidence intervals on the log-hazard scale and back transform */
- /* First for the interval-specific estimates */

```
if se_p ne 0 then do;
```

- /* SE on the log-hazard scale using Taylor series approximation */

```
se_lh_p=sqrt( se_p**2/(p*log(p))**2 );
```

- /* Confidence limits on the log-hazard scale */

```
lo_lh_p=log(-log(p))+1.96*se_lh_p;
hi_lh_p=log(-log(p))-1.96*se_lh_p;
```

SAS Code: Cont…..

- /* Confidence limits on the survival scale (observed survival) */

```
lo_p=exp(-exp(lo_lh_p));
hi_p=exp(-exp(hi_lh_p));
```

- /* Confidence limits for the corresponding relative survival rate */

```
lo_r=lo_p/p_star;
hi_r=hi_p/p_star;
```

- /* Drop temporary variables */

```
drop se_lh_p lo_lh_p hi_lh_p;
```

SAS Code: Cont….

- /* Formats and labels */

```
format lo_p hi_p lo_r hi_r 8.5;
label
lo_p='Lower 95% CI for P'
hi_p='Upper 95% CI for P'
lo_r='Lower 95% CI for R'
hi_r='Upper 95% CI for R'
 ;
end;
```

SAS Code: Cont…

- /* Now for the cumulative estimates */

```
if se_cp ne 0 then do;
```

- /* SE on the log-hazard scale using Taylor series approximation */

```
se_lh_cp=sqrt( se_cp**2/(cp*log(cp))**2 );
```

- /* Confidence limits on the log-hazard scale */

```
lo_lh_cp=log(-log(cp))+1.96*se_lh_cp;
hi_lh_cp=log(-log(cp))-1.96*se_lh_cp;
```

SAS Code: Cont…

- /* Confidence limits on the survival scale (observed survival) */

```
lo_cp=exp(-exp(lo_lh_cp));
hi_cp=exp(-exp(hi_lh_cp));
```

- /* Confidence limits for the corresponding relative survival rate */

```
lo_cr=lo_cp/cp_star;
hi_cr=hi_cp/cp_star;
```

SAS Code: Cont…

```
/* Drop temporary variables */
        drop se_lh_cp lo_lh_cp hi_lh_cp;
  /* Formats and labels */
        format lo_cp hi_cp lo_cr hi_cr 8.5;
    label
        lo_cp='Lower 95% CI for CP'
        hi_cp='Upper 95% CI for CP'
        lo_cr='Lower 95% CI for CR'
        hi_cr='Upper 95% CI for CR'
        ;
    end;
drop se_temp;
```

SAS Code: Cont…

```
label
        range='Interval‘
        fu='Interval'
        l='Alive at start'
        l_prime='Effective number at risk'
        ns='Number surviving the interval'
        d='Deaths'
        w='Withdrawals'
```

SAS Code: Cont..

```
p='Interval-specific observed survival'
cp='Cumulative observed survival'
r='Interval-specific relative survival'
cr='Cumulative relative survival'
p_star='Interval-specific expected survival'
cp_star='Cumulative expected survival'
```

SAS Code: Cont…

```
ln_y_group='ln(l_prime-d/2)'
ln_y='ln(person-time) (using exact times)'
y='Person-time at risk (using exact times)'
d_star='Expected deaths (using exact times)'
d_star_group='Expected deaths (approximate)'
excess='Empirical excess hazard'
```

SAS Code: Cont…

```
        se_p='Standard error of P'
        se_r='Standard error of R'
        se_cp='Standard error of CP'
        se_cr='Standard error of CR'
        ;
run;
```

SAS Code: Cont…

- Print the lifetables. We first need to extract the last variable in the varlist to use as the argument in the pageby command.

```
%let lastvar = %scan(&vars,-1);
```

SAS Code: Print

```
proc print data=&grouped noobs label;
	title2 'Life table estimates of patient survival';
	title3 'The Ederer II method is used to estimate expected survival';
	by &vars;
	pageby &lastvar;
	var range l d w l_prime p cp p_star cp_star r cr;
	format fu 3.0 l d w 4.0 l_prime 8.1 p cp p_star
		cp_star r cr se_p se_r se_cp se_cr 8.5;
	label l='L' d='D' w='W';
run;
```

Final Output

Colon carcinoma diagnosed in Finland 1975-1994 (follow-up to 1995)

Life table estimates of patient

The Ederer II method is used to estimate expected survival

- --------------------------------- Sex=Male Year of diagnosis=1975-84 Age at diagnosis=0-44 -----------------------------

Interval	L	D	W	Effective number at risk	Interval-specific observed survival	Cumulative observed survival	Interval-specific expected survival	Cumulative expected survival	Interval-specific relative survival	Cumulative relative survival
0.0 - 1.0	75	4	0	75.0	0.94667	0.94667	0.99697	0.99697	0.94954	0.94954
1.0 - 2.0	71	8	0	71.0	0.88732	0.84000	0.99682	0.99381	0.89015	0.84524
2.0 - 3.0	63	1	1	62.5	0.98400	0.82656	0.99649	0.99032	0.98747	0.83464
3.0 - 4.0	61	3	0	61.0	0.95082	0.78591	0.99625	0.98660	0.95440	0.79658
4.0 - 5.0	58	3	0	58.0	0.94828	0.74526	0.99601	0.98266	0.95208	0.75841
5.0 - 6.0	55	2	0	55.0	0.96364	0.71816	0.99562	0.97836	0.96787	0.73404
6.0 - 7.0	53	0	0	53.0	1.00000	0.71816	0.99532	0.97378	1.00470	0.73749
7.0 - 8.0	53	0	0	53.0	1.00000	0.71816	0.99491	0.96882	1.00512	0.74127
8.0 - 9.0	53	1	0	53.0	0.98113	0.70461	0.99453	0.96352	0.98653	0.73128
9.0 - 10.0	52	2	0	52.0	0.96154	0.67751	0.99418	0.95792	0.96717	0.70727

Acknowledgement

- This presentation has been adapted from the website of Dr. Paul Dickman (www.pauldickman.com)
- Dr. Paul Dickman,

  Department of Medical Epidemiology and Biostatistics, Karolinska Institutet,

  Stockholm, Sweden
- Riaz Alvi, Saskatchewan Cancer Agency
- Tong Zhu, Saskatchewan Cancer Agency
- Gopinath Narasimhan, Saskatchewan Cancer Agency

Thank you

?