Biostat3: Exercise on risk estimation

Mark Clements

March 15, 2017

Solutions are provided in `Stata`, `R` and `SAS`.

Stata

In outline, we consider various approaches to estimate 5-year risks of colon cancer death for females following a colon cancer diagnosis by stage of disease.

We read in the dataset:

```
. set linesize 80
. cd /home/marcle/repos/biostat3_2014/tmp/
/home/marcle/repos/biostat3_2014/tmp
. use colon, clear
```

We then stset the data (after creating the `id` variable):

```
. capture drop id
. gen id = _n
. stset exit, fail(status==1) origin(dx) id(id) scale(365.24)

                id:  id
     failure event:  status == 1
obs. time interval:  (exit[_n-1], exit]
 exit on or before:  failure
    t for analysis:  (time-origin)/365.24
            origin:  time dx

--------------------------------------------------------------------------------
      15564  total observations
          0  exclusions
--------------------------------------------------------------------------------
      15564  observations remaining, representing
      15564  subjects
       8369  failures in single-failure-per-subject data
  58546.178  total analysis time at risk and under observation
                                                at risk from t =         0
                                     earliest observed entry t =         0
                                          last observed exit t =  20.98073
```

As a first approach, we can calculate non-parametric risk (or failure) estimates from the Kaplan-Meier curves using the `sts list` command:

```
. sts list if agegrp==2 & stage!=0 & sex==2, by(stage) at(5 10) failure

         failure _d:  status == 1
   analysis time _t:  (exit-origin)/365.24
             origin:  time dx
                 id:  id
```

```
            Beg.                        Failure       Std.
    Time     Total     Fail             Function     Error     [95% Conf. Int.]
-------------------------------------------------------------------------------
Localised
       5       793      310              0.2317    0.0117     0.2097    0.2556
      10       338       52              0.2956    0.0138     0.2694    0.3237
Regional
       5       136      225              0.5299    0.0250     0.4819    0.5796
      10        50       18              0.6109    0.0274     0.5577    0.6646
Distant
       5        87     1102              0.9072    0.0088     0.8890    0.9235
      10        39       15              0.9276    0.0084     0.9101    0.9429
-------------------------------------------------------------------------------
Note: Failure function is calculated over full data and evaluated at indicated
      times; it is not calculated from aggregates shown at left.
```

As a second approach, we can fit a Poisson regression with constant rates:

```
. streg i.sex##i.stage if agegrp==2 & stage!=0, dist(exp) nolog base

         failure _d:  status == 1
   analysis time _t:  (exit-origin)/365.24
             origin:  time dx
                 id:  id

Exponential regression -- log relative-hazard form

No. of subjects =        5,735                 Number of obs    =       5,735
No. of failures =        3,087
Time at risk    =  22828.45526
                                               LR chi2(5)       =     4300.57
Log likelihood  =   -8039.8071                 Prob > chi2      =      0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |
        Male |          1  (base)
      Female |   .8576709    .065315    -2.02   0.044     .7387515    .9957332
             |
       stage |
  Localised  |          1  (base)
   Regional  |   2.642185   .2632503     9.75   0.000     2.173477     3.21197
    Distant  |   13.62924   .9005805    39.53   0.000     11.97365    15.51374
             |
   sex#stage |
     Female #|
   Regional  |   1.258471   .1618644     1.79   0.074     .9780514     1.61929
     Female #|
    Distant  |   1.070361   .0947776     0.77   0.443     .8998258    1.273215
             |
       _cons |   .0463934   .0026307   -54.15   0.000     .0415135     .051847
------------------------------------------------------------------------------
```

We can obtain confidence intervals for the rates using the `lincom` command:

```
. lincom _cons+2.sex, eform

 ( 1)  [_t]2.sex + [_t]_cons = 0
```

```
------------------------------------------------------------------------------
          _t |     exp(b)   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   .0397903   .0020227   -63.43   0.000     .0360171    .0439588
------------------------------------------------------------------------------
```

Using the confidence intervals for the rate, we can calculate the five-year risk and its 95% confidence intervals:

```
. di 1-exp(-.0397903*5)
.18041036
. di 1-exp(-.0360171*5)
.1648012
. di 1-exp(-.0439588*5)
.19731587
```

Finally, we can use Cox regression:

```
. stcox i.sex##i.stage if agegrp==2 & stage!=0, nolog base

         failure _d:  status == 1
   analysis time _t:  (exit-origin)/365.24
             origin:  time dx
                 id:  id

Cox regression -- Breslow method for ties

No. of subjects =        5,735                 Number of obs    =       5,735
No. of failures =        3,087
Time at risk    =  22828.45526
                                               LR chi2(5)       =     2775.22
Log likelihood  =    -23864.82                 Prob > chi2      =      0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |
        Male |          1  (base)
      Female |   .9418098   .0717916    -0.79   0.432     .8111074    1.093574
             |
       stage |
  Localised |          1  (base)
   Regional |   2.285848   .2278625     8.29   0.000     1.880166    2.779064
    Distant |   8.842267   .5936013    32.47   0.000     7.752121    10.08572
             |
   sex#stage |
     Female #|
   Regional |   1.179394   .1517318     1.28   0.200     .9165369    1.517637
     Female #|
    Distant |   1.053203   .0932987     0.59   0.558      .885335      1.2529
------------------------------------------------------------------------------
```

We can pull out $S_0(5)$ using the `predict` function with the `basesurv` argument. One way to get the times is the following:

```
. preserve
. sort _t
. predict S0, basesurv
```

```
(9,829 missing values generated)
. keep if agegrp==2 & stage!=0 & _t<5
(11,506 observations deleted)
. list _t S0 if _n==_N

      +----------------------+
      |        _t         S0 |
      |----------------------|
4058. | 4.9802869   .7434324 |
      +----------------------+
. restore
```

We can combine the baseline survival with the hazard ratio to calculate the risk. The risk equation is $Risk = 1 - S_0(t)^{HR}$. Plugging in these estimates for females at five years:

```
. display 1-.74343243^.9418098
.24363054
```

Confidence interval estimation can be done with the use of the `survi` user-contributed command:

```
. survci if agegrp==2 & stage!=0, survival at(sex=2, stage=1) outfile(pred, rep
> lace)
(sex=2.00; stage=1.00)

file pred.dta saved
. preserve
. use pred, clear
(estimates adjusted for sex=2; stage=1)
. sort _t
. keep if _t<5
(1,677 observations deleted)
. gen _fail = 1-_surv
. gen _fail_lb = 1-_ub
. gen _fail_ub = 1-_lb
. list _t _fail _fail_lb _fail_ub if _n==_N

      +-------------------------------------------+
      |        _t      _fail   _fail_lb   _fail_ub |
      |-------------------------------------------|
4058. | 4.9802869   .2436305   .2230176   .2658031 |
      +-------------------------------------------+
. restore
```

R

Initially, load the required packages and read in the data.

```
options(width=80)
require(foreign)
require(survival)
colon <- read.dta("/home/marcle/repos/biostat3_2014/tmp/colon.dta")
nrow(colon)
```

```
[1] 15564
```

For the Kaplan-Meier analysis, we initially calculate the time on study (`tmExit`), subset the dataset, calculate the Kaplan-Meier estimates, and then summarise the risks.

```
require(dplyr)
colon <- mutate(colon, tmExit=as.numeric(exit - dx)/365.24)
tmp <- filter(colon, agegrp=="60-74" & sex=="Female")
fit1 <- survfit(Surv(tmExit,status=="Dead: cancer")~stage, data=tmp)
with(summary(fit1, time=c(5,10)),
     data.frame(strata, time, fail=1-surv, lower=1-upper, upper=1-lower))
## summary(fit1, time=c(5,10))
```

```
           strata time      fail     lower     upper
1   stage=Unknown    5 0.3892831 0.3417862 0.4333527
2   stage=Unknown   10 0.4169903 0.3673982 0.4626946
3 stage=Localised    5 0.2316504 0.2083740 0.2542424
4 stage=Localised   10 0.2955891 0.2679427 0.3221915
5  stage=Regional    5 0.5299454 0.4783378 0.5764476
6  stage=Regional   10 0.6109463 0.5534227 0.6610603
7   stage=Distant    5 0.9071782 0.8882195 0.9229214
8   stage=Distant   10 0.9276370 0.9092668 0.9422878
```

We now re-do this analysis using Poisson regression and assuming a constant rate across time.

```
## streg i.sex##i.stage if agegrp==2 & stage!=0, dist(exp) nolog base
tmp <- filter(colon, agegrp=="60-74" & stage!="Unknown")
fit2 <- glm(I(status=="Dead: cancer")~factor(sex)*factor(stage)+
offset(log(tmExit)), data=tmp, family=poisson)
summary(fit2)
```

```
Call:
glm(formula = I(status == "Dead: cancer") ~ factor(sex) * factor(stage) +
    offset(log(tmExit)), family = poisson, data = tmp)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.8782  -0.7964  -0.1320   1.3201   3.2919

Coefficients:
                                      Estimate Std. Error z value Pr(>|z|)
(Intercept)                           -3.07060    0.05670 -54.151   <2e-16
factor(sex)Female                     -0.15353    0.07615  -2.016   0.0438
factor(stage)Regional                  0.97161    0.09963   9.752   <2e-16
factor(stage)Distant                   2.61222    0.06608  39.533   <2e-16
factor(sex)Female:factor(stage)Regional  0.22990    0.12862   1.787   0.0739
factor(sex)Female:factor(stage)Distant   0.06800    0.08855   0.768   0.4425

(Intercept)                           ***
factor(sex)Female                     *
factor(stage)Regional                 ***
factor(stage)Distant                  ***
factor(sex)Female:factor(stage)Regional .
factor(sex)Female:factor(stage)Distant
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 14206.2  on 5734  degrees of freedom
Residual deviance:  9905.6  on 5729  degrees of freedom
AIC: 16092

Number of Fisher Scoring iterations: 7
```

We use the `predict` function to calculate the confidence interval for the log-rates, then calculate the rates and the risks.

```
pred2 <- predict(fit2, type="link",
 newdata=data.frame(sex="Female",stage="Localised",tmExit=1),
 se.fit=TRUE)
rates <- with(pred2,
      exp(cbind(fit, lower=fit-1.96*se.fit, upper=fit+1.96*se.fit)))
1-exp(-5*rates)
```

```
       fit     lower     upper
1 0.1804104 0.1648007 0.1973163
```

Finally, we fit a Cox regression using Breslow's methods for ties for comparison with Stata.

```
tmp <- filter(colon, agegrp=="60-74" & stage!="Unknown")
fit3 <- coxph(Surv(tmExit, status=="Dead: cancer")~factor(sex)*factor(stage),
      data=tmp, ties="breslow")
summary(fit3)
```

```
Call:
coxph(formula = Surv(tmExit, status == "Dead: cancer") ~ factor(sex) *
    factor(stage), data = tmp, ties = "breslow")

  n= 5735, number of events= 3087

                                        coef exp(coef) se(coef)      z
factor(sex)Female                    -0.05995   0.94181  0.07623 -0.786
factor(stage)Regional                 0.82674   2.28585  0.09968  8.294
factor(stage)Distant                  2.17954   8.84227  0.06713 32.466
factor(sex)Female:factor(stage)Regional  0.16500   1.17939  0.12865  1.283
factor(sex)Female:factor(stage)Distant   0.05184   1.05320  0.08859  0.585
                                     Pr(>|z|)
factor(sex)Female                       0.432
factor(stage)Regional                  <2e-16 ***
factor(stage)Distant                   <2e-16 ***
factor(sex)Female:factor(stage)Regional    0.200
factor(sex)Female:factor(stage)Distant     0.558
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

                                     exp(coef) exp(-coef) lower .95
factor(sex)Female                       0.9418     1.0618    0.8111
factor(stage)Regional                   2.2858     0.4375    1.8802
factor(stage)Distant                    8.8423     0.1131    7.7521
factor(sex)Female:factor(stage)Regional    1.1794     0.8479    0.9165
factor(sex)Female:factor(stage)Distant     1.0532     0.9495    0.8853
                                     upper .95
factor(sex)Female                        1.094
factor(stage)Regional                    2.779
factor(stage)Distant                    10.086
factor(sex)Female:factor(stage)Regional     1.518
factor(sex)Female:factor(stage)Distant      1.253

Concordance= 0.753  (se = 0.005 )
Rsquare= 0.384   (max possible= 1 )
Likelihood ratio test= 2775  on 5 df,   p=0
Wald test            = 2480  on 5 df,   p=0
Score (logrank) test = 3259  on 5 df,   p=0
```

To calculate the predicted risks, we `survfit` the `coxph` model for given covariates, use `summary` for the specific times, and calculate the risks. Interestingly, these estimates are not identical to Stata, although the Cox fit is very similar.

```
pred3 <- survfit(fit3, newdata=data.frame(sex="Female",stage="Localised"))
with(summary(pred3, time=c(5,10)),
     data.frame(time, fail=1-surv, lower=1-upper, upper=1-lower))
```

```
  time      fail     lower     upper
1    5 0.2420369 0.2203020 0.2631659
2   10 0.2881452 0.2630863 0.3123519
```

SAS

Initially, set the environment and read in the data. The `%reset` macro is specific to my set-up, so this can be ignored.

```
1982  options linesize=80;
1983  %macro reset;
1984      options obs=max nosyntaxcheck replace;
1985  %mend reset;
1986  %reset;
NOTE: The PROCEDURE SQL printed page 303.
NOTE: PROCEDURE SQL used (Total process time):
      real time           13.42 seconds
      cpu time            0.01 seconds

1987  proc import datafile="/home/marcle/repos/biostat3_2014/tmp/colon.dta"
1987! out=colon replace;
1988  run;
NOTE: The import data set has 15564 observations and 13 variables.
NOTE: WORK.COLON data set was successfully created.
NOTE: PROCEDURE IMPORT used (Total process time):
      real time           0.00 seconds
      cpu time            0.01 seconds
```

Now, calculate the follow-up time and calculate the Kaplan-Meier estimators. I like to use SQL for data manipulation – one could do this differently using data steps and merging.

```
1990  data colon;
1991      set colon;
1992      tmexit = (exit-dx)/365.24;
1993  run;
NOTE: There were 15564 observations read from the data set WORK.COLON.
NOTE: The data set WORK.COLON has 15564 observations and 14 variables.
NOTE: DATA statement used (Total process time):
      real time           0.00 seconds
      cpu time            0.00 seconds

1994  proc lifetest data=colon notable outsurv=surv;
1995      time tmexit*status(0,2,4);
1996      strata stage;
1997      where agegrp=2 and sex=2;
1998  run;
NOTE: The LOGLOG transform is used to compute the confidence limits for the
      survivor function. To suppress using this transform, specify
      CONFTYPE=LINEAR in the PROC LIFETEST statement.
                                The SAS System                          304
                                                  16:25 Wednesday, March 15, 2017
```

The LIFETEST Procedure

Summary of the Number of Censored and Uncensored Values

Stratum	stage	Total	Failed	Censored	Percent Censored
1	Distant	1254	1123	131	10.45
2	Localised	1488	387	1101	73.99
3	Regional	465	248	217	46.67
4	Unknown	487	190	297	60.99
Total		3694	1948	1746	47.27

The LIFETEST Procedure

Testing Homogeneity of Survival Curves for tmexit over Strata

Rank Statistics

stage	Log-Rank	Wilcoxon
Distant	735.62	2045502
Localised	-630.44	-1706540
Regional	-11.76	-124435
Unknown	-93.42	-214527

Covariance Matrix for the Log-Rank Statistics

stage	Distant	Localised	Regional	Unknown
Distant	291.196	-185.943	-51.959	-53.294
Localised	-185.943	468.352	-133.362	-149.047
Regional	-51.959	-133.362	222.430	-37.109
Unknown	-53.294	-149.047	-37.109	239.451

Covariance Matrix for the Wilcoxon Statistics

stage	Distant	Localised	Regional	Unknown
Distant	2.4668E9	-1.549E9	-4.606E8	-4.57E8
Localised	-1.549E9	3.3774E9	-9.07E8	-9.211E8
Regional	-4.606E8	-9.07E8	1.6281E9	-2.604E8
Unknown	-4.57E8	-9.211E8	-2.604E8	1.6385E9

Test of Equality over Strata

Test	Chi-Square	DF	Pr > Chi-Square
Log-Rank	1945.9735	3	<.0001

```
                  Wilcoxon   1772.5785       3      <.0001
                  -2Log(LR)  2530.4314       3      <.0001
NOTE: The data set WORK.SURV has 2380 observations and 7 variables.
NOTE: The PROCEDURE LIFETEST printed pages 304-305.
NOTE: PROCEDURE LIFETEST used (Total process time):
      real time           0.00 seconds
      cpu time            0.00 seconds

1999  proc sql noerrorstop;
2000      select distinct stage, tmexit, 1-survival as failure, 1-sdf_ucl as
2000! fail_lcl, 1-sdf_lcl as fail_ucl
2001          from surv
2002          natural join (select stage, max(tmexit) as tmexit from surv where
2002! tmexit<5 and _censor_=0 group by stage)
2003          where _censor_=0
2004          order by stage;
                                The SAS System                              306
                                                  16:25 Wednesday, March 15, 2017

                   stage    tmexit   failure  fail_lcl  fail_ucl
                 ------------------------------------------------
                       0  4.963859  0.389283  0.345338  0.436735
                       1  4.958384   0.23165  0.209662  0.255551
                       2  4.709232  0.529945  0.481908  0.579625
                       3   4.87077  0.907178  0.889008  0.923494
2005  quit;
NOTE: The PROCEDURE SQL printed page 306.
79                              The SAS System  16:25 Wednesday, March 15, 2017

NOTE: PROCEDURE SQL used (Total process time):
      real time           0.00 seconds
      cpu time            0.01 seconds
```

Now, fit the Poisson regression.

```
2007  data colon;
2008      set colon;
2009      event = (status=1);
2010      logpt=log(tmExit);
2011  run;
NOTE: There were 15564 observations read from the data set WORK.COLON.
NOTE: The data set WORK.COLON has 15564 observations and 16 variables.
NOTE: DATA statement used (Total process time):
      real time           0.00 seconds
      cpu time            0.01 seconds

2012  ods output estimates=est;
2013  proc genmod data=colon;
2014      class sex stage(ref='Localised');
2015      model event = sex|stage / dist=poisson offset=logpt;
2016      estimate "Rate"
2017          intercept 1
2018          sex 1 0
2019          stage 0 0 1
2020          sex*stage 0 0 1 0 0 0;
2021      where agegrp=2 and stage~=0;
2022  run;
                                The SAS System                              307
                                                  16:25 Wednesday, March 15, 2017
```

The GENMOD Procedure

Model Information

Data Set	WORK.COLON
Distribution	Poisson
Link Function	Log
Dependent Variable	event
Offset Variable	logpt

Number of Observations Read	5735
Number of Observations Used	5735

Class Level Information

Class	Levels	Values
sex	2	Female Male
stage	3	Distant Regional Localised

Parameter Information

Parameter	Effect	sex	stage
Prm1	Intercept		
Prm2	sex	Female	
Prm3	sex	Male	
Prm4	stage		Distant
Prm5	stage		Regional
Prm6	stage		Localised
Prm7	sex*stage	Female	Distant
Prm8	sex*stage	Female	Regional
Prm9	sex*stage	Female	Localised
Prm10	sex*stage	Male	Distant
Prm11	sex*stage	Male	Regional
Prm12	sex*stage	Male	Localised

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	5729	9905.6142	1.7290
Scaled Deviance	5729	9905.6142	1.7290
Pearson Chi-Square	5729	52631.0454	9.1868
Scaled Pearson X2	5729	52631.0454	9.1868
Log Likelihood		-8039.8071	
Full Log Likelihood		-8039.8071	
AIC (smaller is better)		16091.6142	
AICC (smaller is better)		16091.6289	
BIC (smaller is better)		16131.5403	

Algorithm converged.
NOTE: Algorithm converged.

Analysis Of Maximum Likelihood Parameter Estimates

Parameter			DF	Estimate	Standard Error	Wald 95% Confidence Limits	
Intercept			1	-3.0706	0.0567	-3.1817	-2.9595
sex	Female		1	-0.1535	0.0762	-0.3028	-0.0043
sex	Male		0	0.0000	0.0000	0.0000	0.0000
stage	Distant		1	2.6122	0.0661	2.4827	2.7417
stage	Regional		1	0.9716	0.0996	0.7763	1.1669
stage	Localised		0	0.0000	0.0000	0.0000	0.0000
sex*stage	Female	Distant	1	0.0680	0.0885	-0.1056	0.2415
sex*stage	Female	Regional	1	0.2299	0.1286	-0.0222	0.4820
sex*stage	Female	Localised	0	0.0000	0.0000	0.0000	0.0000
sex*stage	Male	Distant	0	0.0000	0.0000	0.0000	0.0000
sex*stage	Male	Regional	0	0.0000	0.0000	0.0000	0.0000
sex*stage	Male	Localised	0	0.0000	0.0000	0.0000	0.0000
Scale			0	1.0000	0.0000	1.0000	1.0000

Analysis Of Maximum Likelihood Parameter Estimates

Parameter			Wald Chi-Square	Pr > ChiSq
Intercept			2932.28	<.0001
sex	Female		4.06	0.0438
sex	Male		.	.
stage	Distant		1562.85	<.0001
stage	Regional		95.10	<.0001
stage	Localised		.	.
sex*stage	Female	Distant	0.59	0.4425
sex*stage	Female	Regional	3.19	0.0739
sex*stage	Female	Localised	.	.
sex*stage	Male	Distant	.	.
sex*stage	Male	Regional	.	.
sex*stage	Male	Localised	.	.
Scale				

NOTE: The scale parameter was held fixed.
NOTE: The scale parameter was held fixed.

Contrast Estimate Results

Label	Mean Estimate	Mean Confidence Limits		L'Beta Estimate	Standard Error	Alpha
Rate	0.0398	0.0360	0.0440	-3.2241	0.0508	0.05

```
                              The GENMOD Procedure

                          Contrast Estimate Results

                              L'Beta             Chi-
            Label      Confidence Limits     Square    Pr > ChiSq

            Rate      -3.3238     -3.1245    4022.9        <.0001
NOTE: The data set WORK.EST has 1 observations and 11 variables.
NOTE: The PROCEDURE GENMOD printed pages 307-309.
NOTE: PROCEDURE GENMOD used (Total process time):
      real time           0.02 seconds
      cpu time            0.02 seconds

2023  proc sql noerrorstop;
2024      select label, 1-exp(-MeanEstimate*5) as risk, 1-exp(-MeanLowerCL*5) as
2024!  lower, 1-exp(-MeanUpperCL*5) as upper from est;
                                The SAS System                              310
                                                  16:25 Wednesday, March 15, 2017

                     Label      risk     lower     upper
                     -----------------------------------
                     Rate    0.18041  0.164801  0.197316
2025  quit;
NOTE: The PROCEDURE SQL printed page 310.
NOTE: PROCEDURE SQL used (Total process time):
      real time           0.00 seconds
      cpu time            0.00 seconds
```

Finally, fit the Cox regression.

```
2027  data colon_subset;
2028      sex=2; agegrp=2;
2029      do stage=1 to 3;
2030          output;
2031          end;
2032      format sex sexa. stage stagea. agegrp agegrpa.;
2033  run;
NOTE: The data set WORK.COLON_SUBSET has 3 observations and 3 variables.
NOTE: DATA statement used (Total process time):
      real time           0.00 seconds
      cpu time            0.01 seconds
80                              The SAS System  16:25 Wednesday, March 15, 2017


2034  proc phreg data=colon;
2035      class sex stage(ref='Localised');
2036      model tmExit*status(0,2,4)=sex|stage;
2037      baseline covariates=colon_subset out=pred3 survival=_all_;
2038      where agegrp=2 and stage~=0;
2039  run;
                                The SAS System                              311
                                                  16:25 Wednesday, March 15, 2017

                              The PHREG Procedure

                               Model Information

      Data Set                 WORK.COLON
```

Dependent Variable	tmexit	
Censoring Variable	status	Vital status at last contact
Censoring Value(s)	0 2 4	
Ties Handling	BRESLOW	

Number of Observations Read	5735
Number of Observations Used	5735

Class Level Information

Class	Value	Design Variables	
sex	Female	1	
	Male	0	
stage	Distant	1	0
	Localised	0	0
	Regional	0	1

Summary of the Number of Event and Censored Values

Total	Event	Censored	Percent Censored
5735	3087	2648	46.17

Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.
NOTE: Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Without Covariates	With Covariates
-2 LOG L	50504.854	47729.639
AIC	50504.854	47739.639
SBC	50504.854	47769.814

The PHREG Procedure

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	2775.2151	5	<.0001
Score	3259.1938	5	<.0001

```
Wald                  2479.6926        5         <.0001
```

Type 3 Tests

Effect	DF	Wald Chi-Square	Pr > ChiSq
sex	1	0.6186	0.4316
stage	2	1126.6664	<.0001
sex*stage	2	1.6516	0.4379

Analysis of Maximum Likelihood Estimates

Parameter			DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq
sex	Female		1	-0.05995	0.07623	0.6186	0.4316
stage	Distant		1	2.17954	0.06713	1054.0599	<.0001
stage	Regional		1	0.82678	0.09968	68.7919	<.0001
sex*stage	Female	Distant	1	0.05184	0.08859	0.3424	0.5584
sex*stage	Female	Regional	1	0.16502	0.12865	1.6453	0.1996

Analysis of Maximum Likelihood Estimates

Parameter			Hazard Ratio
sex	Female		.
stage	Distant		.
stage	Regional		.
sex*stage	Female	Distant	.
sex*stage	Female	Regional	.

Analysis of Maximum Likelihood Estimates

Parameter			Label
sex	Female		Sex Female
stage	Distant		Clinical stage at diagnosis Distant
stage	Regional		Clinical stage at diagnosis Regional
sex*stage	Female	Distant	Sex Female * Clinical stage at diagnosis Distant
sex*stage	Female	Regional	Sex Female * Clinical stage at diagnosis Regional

```
NOTE: Hazard ratios that cannot be conveniently calculated or displayed are set
      to missing in the ParameterEstimates table. Use the HAZARDRATIO statement
      to compute the needed hazard ratios.
NOTE: The data set WORK.PRED3 has 1065 observations and 8 variables.
NOTE: The PROCEDURE PHREG printed pages 311-312.
NOTE: PROCEDURE PHREG used (Total process time):
      real time           0.01 seconds
      cpu time            0.01 seconds

2040  proc sql noerrorstop;
2041      select distinct stage, tmexit, 1-survival as failure, 1-UpperSurvival
2041! as fail_lcl, 1-LowerSurvival as fail_ucl
2042          from pred3
2043          natural join (select stage, max(tmexit) as tmexit from pred3 where
```

```
2043!  tmexit<5 group by stage)
2044           where sex=2 and agegrp=2
2045           order by stage;
                                The SAS System                                 313
                                                    16:25 Wednesday, March 15, 2017

                    stage    tmexit   failure  fail_lcl  fail_ucl
                   ------------------------------------------------
                        1  4.958384  0.242035    0.2203  0.263164
                        2  4.958384  0.526277   0.47947  0.568875
                        3  4.958384   0.92428  0.910237  0.936127
```

I always forget that SAS output is so verbose. Questions or revisions: please direct to me (mark.clements@ki.se).