STATS 210 SAS LAB TEN, August 9, 2004

Lab Ten: Survival Analysis II: Cox Regression

Lab Objectives

After today's lab you should be able to:

1. Practice dealing with SAS date/time variables.
2. Fit models using PROC PHREG. Understand PROC PHREG output.
3. Understand output from the "baseline" statement.
4. Output estimated survivor functions and plot cumulative hazards.
5. Understand the role of the strata statement in PROC PHREG.
6. Evaluate PH assumption graphically.
7. Output and plot predicted survivor functions at user-specified levels of the covariates.

STATS 210 SAS LAB TEN, August 9, 2004

Follow along with the computer in front…

1. Double-click on to open the SAS editor file "data creation code" which should be saved in your stats210 folder from last week; run the libname statement:

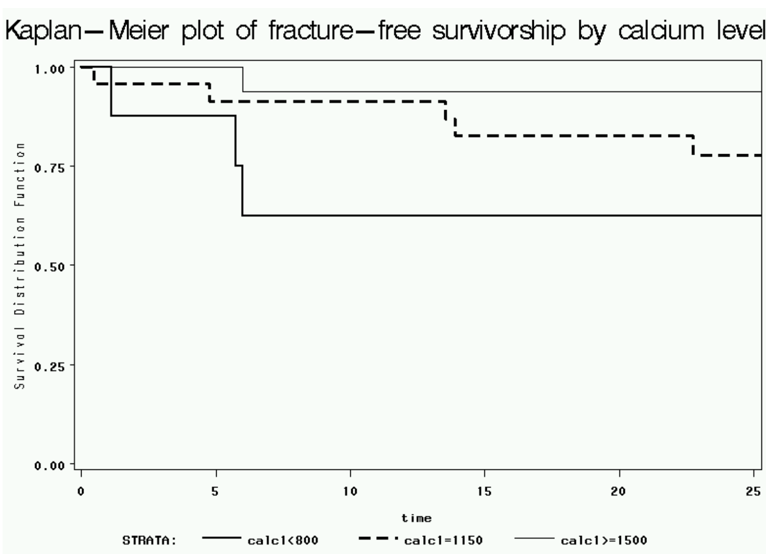
```
libname stats210 'C:\Documents and Settings\mitl-pc.LANE-LIB\My Documents\Stats210';
```

2. Using the Explorer Browser on the left hand side of your screen, double check that a stats210 library has been properly created, and that it contains the SAS dataset "runners"

3. Plot the Kaplan-Meier survival curve for the by level of daily calcium intake (<800 mg/day, 800-1499 mg/day, and 1500+ mg day) using the strata statement as below:

```
proc lifetest data=stats210.runners plots=(s) graphics censoredsymbol=none
maxtime=24;
	time time*sf1(0);
	strata calc1(800,1500);
	title 'Kaplan-Meier plot of fracture-free survivorship by calcium
	level';
	symbol1 v=none c=black w=2 i=join line=1;
	symbol2 v=none c=black w=2 i=join line=2;
	symbol3 v=none c=black w=2 i=join line=3;
run;
```

This asks SAS to divide into groups as follows: [-∞ ,800) [800, 1500) [1500, ∞).
This is an extremely useful feature of PROC LIFETEST, because you don't have to break the variables into categories yourself; SAS does it for you.



With small numbers in each group, this is a pretty striking difference in fracture rates by calcium level, though it is not statistically significant yet (p=.13). A significant log-rank test here would mean that at least one pair of groups is significantly different in fracture rate.

Test of Equality over Strata

Test	Chi-Square	DF	Pr > Chi-Square
Log-Rank	**3.9352**	**2**	**0.1398**
Wilcoxon	4.0947	2	0.1291
-2Log(LR)	4.5365	2	0.1035

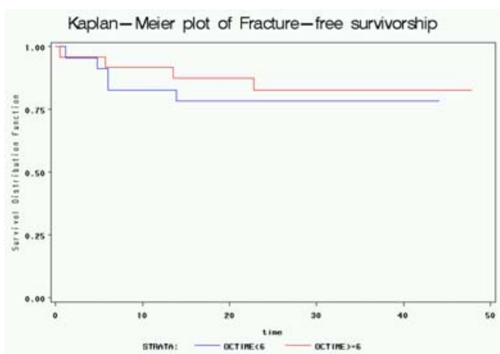
4. Last time, we saw that there was no difference in fracture rates between those randomized to control and those randomized to treatment (so called "intention-to-treat analysis"). We might also want to know if there is a difference in fracture rates between those who actually took oral contraceptives and those who did not. We recorded the start and stop dates of taking treatment in the date variables "startoc" and "endoc." Use these variables to create a new variable that represents the amount of time (in months) that each woman actually took the treatment (regardless of her assignment to treatment or control). Recall that date variables in SAS are recorded as the number of days since Jan 1., 1960.

```
data stats210.runners;
set stats210.runners;
OCTIME=(endoc-startoc)/365.25*12;
run;
```

5. Use: Solutions→Analysis→Interactive Data Analysis→Stats210→Runners→OCTIME→Analyze→Distribution→Tables→Frequency counts, to examine the distribution of time-on-treatment for the study participants. Notice it's basically bimodal—women either took OCs for 0 or very few months or for more than 1 year.

6. Plot a Kaplan-Meier Curve that compares women who took OC's for at least 6 months against women who never took them:

```
proc lifetest data=stats210.runners plots=(s) graphics
censoredsymbol=none;
     time time*sf1(0);
     title 'Kaplan-Meier plot of Fracture-free survivorship';
     strata octime (6);
run;
```



7. Use Cox Regression to examine the relationship between treatment use (actual) and treatment randomization (intention-to-treat) and generate hazard ratios with confidence limits:

Syntax for PHREG

```
proc phreg data=stats210.runners;
     model time*sf1(0)=octime / risklimits;
     title 'Cox model for runners data—actual OC time';
run;

proc phreg data=stats210.runners;
     model time*sf1(0)=treatr / risklimits;
     title 'Cox model for runners data—treatment randomization';
run;
```

Asks for 95% confidence limits for the hazard ratios.

Examine Output:

```
     Cox model for runners data-actual OC time
                                  The PHREG Procedure

                                  Model Information

                    Data Set                 STATS210.RUNNERS
                    Dependent Variable       time                  time
                    Censoring Variable       sf1                   SF1
                    Censoring Value(s)       0
                    Ties Handling            BRESLOW

                    Summary of the Number of Event and Censored Values

                                                            Percent
                          Total       Event    Censored    Censored

                             47           9          38       80.85


                                  Convergence Status

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


                                 Model Fit Statistics

                                        Without           With
                          Criterion     Covariates     Covariates

                          -2 LOG L          67.500         67.444
                          AIC               67.500         69.444
                          SBC               67.500         69.641

                          Testing Global Null Hypothesis: BETA=0

                  Test                   Chi-Square       DF     Pr > ChiSq

                  Likelihood Ratio           0.0555        1         0.8138
                  Score                      0.0552        1         0.8142
                  Wald                       0.0551        1         0.8144


                          Analysis of Maximum Likelihood Estimates

                 Parameter    Standard                                Hazard   95% Hazard Ratio
 Variable  DF     Estimate       Error  Chi-Square  Pr > ChiSq         Ratio   Confidence Limits

 OCTIME     1     -0.00651     0.02772      0.0551      0.8144         0.994     0.941     1.049
```

Method used for dealing with ties in event times; Breslow is default

Test the global null hypothesis that all coefficients are equal to 0.

0.6% decrease in hazard rate (=instantaneous risk of fracture) for every 1-month increase in OC use. Very close to the null value.

For treatment randomization:

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	0.0193	1	0.8895
Score	0.0194	1	0.8893
Wald	0.0194	1	0.8893

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	95% Hazard Ratio Confidence Limits		Variable Label
treatr	1	0.09336	0.67097	0.0194	0.8893	**1.098**	**0.295**	**4.090**	**TREATR**

9% increase in hazard for being randomized to treatment. Also close to the null value, and similar to Odds Ratio results from logistic regression and contingency tables.

8. There doesn't seem to be much there. But we might think that the relationship between actual time on OC's and fracture protection is not linear. Could try OC use as a binary predictor (as in step 5). Note the convenient ability to create new (temporary) variables within PROC PHREG:

```
proc phreg data=stats210.runners;
     model time*sf1(0)=oc / risklimits;
     if octime>=6 then oc=6 ;
     if octime<6 then oc=0 ;
     title 'Cox model for runners data-OC use >=6 month';
run;
```

Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	95% Hazard Ratio Confidence Limits	
oc	1	-0.04785	0.11183	0.1831	0.6687	**0.953**	**0.766**	**1.187**

Essentially no difference.

9. Another convenient feature of PROC PHREG is that you can allow that the baseline hazard might be different across different groups, and you can stratify on different groups. For example, in the runners study there is reason to believe that the athletes recruited from different sites might have very different characteristics (some sites recruited mostly from collegiate runners and others recruited only from post-collegiate runners) and might have very different baseline fracture rates. Therefore, we can stratify on fracture rates as follows:

```
proc phreg data=stats210.runners;
     model time*sf1(0)=oc / risklimits;
     if octime>=6 then oc=6 ;
     if octime<6 then oc=0 ;
     strata sitenum;
     title 'Cox model for runners data-OC use >=6 month';
run;
```

10. Evaluate the effects of several predictors, each adjusted for each other:

```
title ' ';
proc phreg data=stats210.runners;
     model time*sf1(0)=calc1 stressf bmc1 edever menarch / risklimits;
     baseline out=outdata survival=S;
run;
proc print data=outdata;
run;
```

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	17.5738	5	0.0035
Score	18.3211	5	0.0026
Wald	11.3943	5	0.0441

Indicates that at least one of the coefficients in the model is not equal to 0.

Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	95% Hazard Ratio Confidence Limits		Variable Label
calc1	1	-0.00172	0.0007759	4.8936	0.0270	0.998	0.997	1.000	calc1
stressf	1	1.90860	0.99539	3.6766	0.0552	6.744	0.959	47.443	STRESSF
bmc1	1	-0.00118	0.00109	1.1765	0.2781	0.999	0.997	1.001	bmc1
edever	1	5.64477	2.31313	5.9552	0.0147	282.809	3.038	26328.80	edever
menarch	1	-0.64796	0.34077	3.6155	0.0572	0.523	0.268	1.020	MENARCH

Controlling for other variables in the model, calcium significantly decreases fracture risk by more than 80% for every 1000mg/day:

$$\beta_{calc} = -.00172$$

$$\beta_{calc}(1000mg) = -.00172(1000) = -1.72$$

$$HR_{1000mg} = e^{-1.72} = .179$$

Controlling for other variables in the model, prior eating disorder increases hazard an estimated 282-fold, BUT the confidence interval ranges from 3-26,328, indicating a terribly large range of possibilities for the true hazard ratio.

Controlling for other variables in the model, prior fracture increases hazard of fracture 6.7-fold.

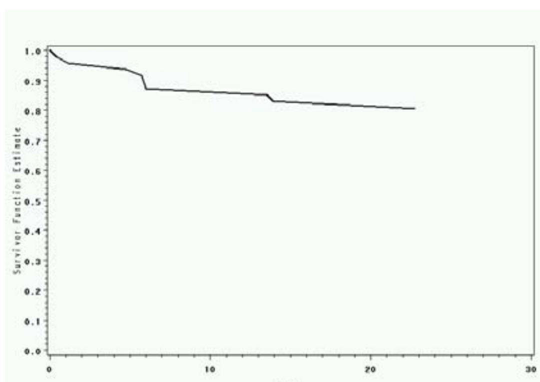
11. SAS can also calculate a baseline survivor function (recall that the baseline survivor function is NOT estimated by Cox regression). SAS uses a non-parametric method to estimate the baseline survivor function. To get the estimated survivor function not accounting for any covariates (similar to Kaplan-Meier):

```
title ' ';

     model time*sf1(0)=/ risklimits;
     baseline out=outdata survival=S;
run;
proc print data=outdata;
run;
axis1 order=(0 to 1.0 by .10) label=(angle=90);
proc gplot data=outdata;
     plot S*time /vaxis=axis1;
     symbol1 i=join c=black line=1;
     symbol2 i=join c=black line=2;
run;
```

Obs	time	S
1	0.0000	1.00000
2	0.4600	0.97872
3	1.1170	0.95745
4	4.7639	0.93617
5	5.7495	0.91489
6	6.0000	0.87234
7	13.5359	0.85106
8	13.8973	0.82979
9	22.7351	0.80608

Estimated cumulative baseline probability of stress-fracture free "survival" at each event time.



Similar to K-M curve, but smooth!

12. You can also use this technique to get predicted curves for individuals at particular values of the predictors. For example, if you want to get the predicted survival function for a woman who has a low calcium intake and low bone mineral density:

```
data mycovs;
     input calc1 bmc1;
     datalines;
     800 1900
     ;
run;
```

Enter covariates into a separate dataset.

```
proc phreg data=stats210.runners;
     model time*sf1(0)=calc1 bmc1  / risklimits;
     baseline out=outdata covariates=mycovs survival=S;
run;
proc print data=outdata;
run;
proc gplot data=outdata;
     plot S*time=calc1 /vaxis=axis1 nolegend;
     symbol1 i=join c=black line=1;
     symbol2 i=join c=black line=2;
run;
```

Then tell SAS to use those covariates to calculate the estimated survival curve

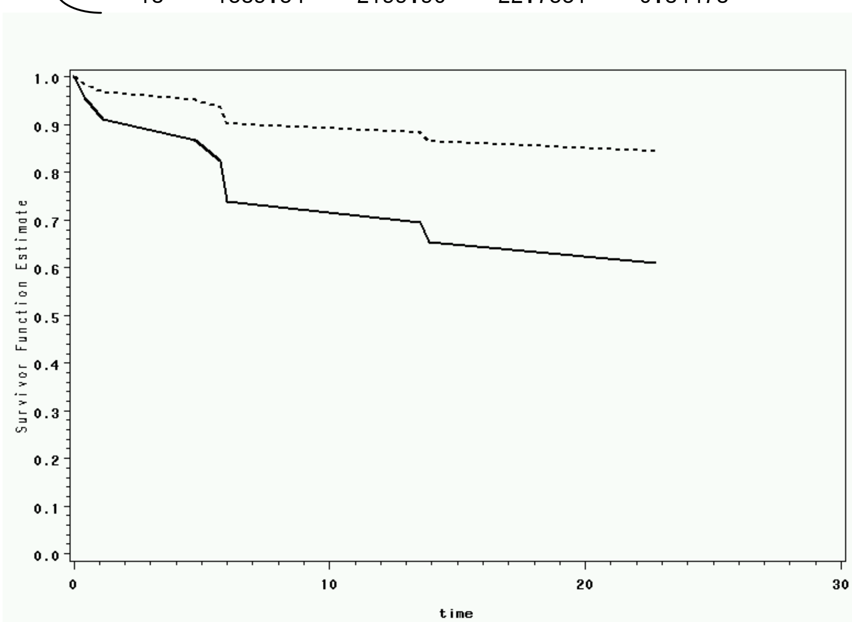
As calc1 and bmc1 distinguish the two sets of survival values (See below), I can use either one to ask for separate survival plots for a woman with the covariates I've specified vs. a woman with the mean values of the covariates (which SAS automatically calculates).

Obs	calc1	bmc1	time	S
1	800.00	1900.00	0.0000	1.00000
2	800.00	1900.00	0.4600	0.95506
3	800.00	1900.00	1.1170	0.91096
4	800.00	1900.00	4.7639	0.86738
5	800.00	1900.00	5.7495	0.82431
6	800.00	1900.00	6.0000	0.73823
7	800.00	1900.00	13.5359	0.69563
8	800.00	1900.00	13.8973	0.65431
9	800.00	1900.00	22.7351	0.60964
10	1389.34	2199.90	0.0000	1.00000
11	1389.34	2199.90	0.4600	0.98445
12	1389.34	2199.90	1.1170	0.96871
13	1389.34	2199.90	4.7639	0.95266
14	1389.34	2199.90	5.7495	0.93627
15	1389.34	2199.90	6.0000	0.90173
16	1389.34	2199.90	13.5359	0.88364
17	1389.34	2199.90	13.8973	0.86539
18	1389.34	2199.90	22.7351	0.84478

Defines survivor curve for woman with calcium intake of 800 mg/day and bone mineral content at baseline of 1900 g.

Defines survivor curve for woman with the sample mean values of calcium and bmc.

Use the option "/ nomean" in the baseline statement of PROC PHREG to suppress estimation of survival function at the mean values of the covariates.

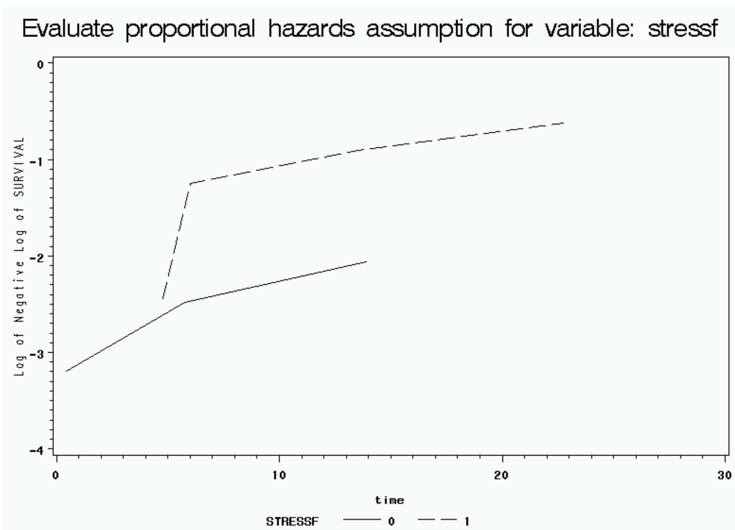


13. We can also use these plots to assess the validity of the proportional hazards assumption.

Note that stressf has been removed from the model statement; stratifying by stressf allows SAS to assume different baseline hazards for each stressf group (which can later be compared to test PH assumption)…

```
axis1 label=(angle=90);
proc phreg data=stats210.runners;
      model time*sf1(0)= ;
      strata stressf;
      baseline out=outdata loglogs=lls;
run;

proc gplot data=outdata;
title 'Evaluate proportional hazards assumption for variable: stressf';
      plot lls*time=stressf /vaxis=axis1;
      symbol1 i=join c=black line=1;
      symbol2 i=join c=black line=2;
run;
```



Similar plot can also be made in LIFETEST by asking for the lls plot:

```
proc lifetest
data=stats210.runners
plots=(lls) graphics
censoredsymbol=none;
time time*sf1(0);
strata stressf;
run;
```