Session 4: Predicted survival

Today we will familiarize ourselves with more of the capabilities of PROC PHREG.

1. Predicted survival

   Let's start by running the nursing home data set and ultimately produce the predicted survival from the PH regression model.

   First we generate the nursing home data set, which we read in from a text file as follows:

```
proc format;
     value marfmt 0='Single' 1='Married';
run;

data nurshome;
     infile 'nurshome.dat';
     input los   age   rx   gender   married   health   fail;
     label los='Length of stay'
           rx='Treatment'
           married='Marriage status'
           health='Health index'
           fail='Censoring index';
     format married marfmt.;
run;
```

   Note the format statement and the rest of the data-step statements.

   The PROC PHREG statements are as follows:

```
proc phreg data=nurshome;
     model los*fail(0)=married health;
     output out=outsurv survival=predsurv;
     title 'PH regression analysis of nursing home data';
run;
```

   Note the new statement

```
     output out=outsurv survival=predsurv;
```

   This statement produces a data set named `outsurv`, which includes, beyond the variables `married` and `health`, the variables `los` and `fail` and the predicted survival at each value of `los`, `presurv`.

The relevant part of the output from the PHREG procedure is as follows:

```
                PH regression analysis of nursing home data                612
                                                21:06 Sunday, February 26, 2012

                           The PHREG Procedure

                  Analysis of Maximum Likelihood Estimates

                        Parameter      Standard
   Parameter    DF      Estimate         Error    Chi-Square    Pr > ChiSq

   married       1       0.29704       0.07219       16.9330        <.0001
   health        1       0.16545       0.03124       28.0467        <.0001

                  Analysis of Maximum Likelihood Estimates

                                    Hazard
                  Parameter          Ratio    Label

                  married            1.346    Marriage status
                  health             1.180    Health index
```

The above output means that $\hat{\beta}_1 = 0.297$ and $\hat{\beta}_2 = 0.165$ corresponding to marital status and the health index respectively.

We would like to sort the data by combination group of `married` and `health`. If we would like to maintain the original, unsorted, data set, then we output the sorted data set with a different name. This we accomplish as follows:

```
proc sort data=outsurv out=prsurvsort;
     by married health los;
run;
```

The statement that produces a new data set `prsurvsort` is

```
proc sort data=outsurv out=prsurvsort;
```

(Otherwise the sorted data set will be stored in the overwritten `outsurv` data set). The printout of the new, sorted, data set is as follows (note that sorting is accomplished from right to left, with `los` being sorted within `health` and the latter within `married`).

```
proc print label data=prsurvsort label;;
     var married health los predsurv;
     title 'Printout of the predicted survival from nursing home data';
run;
```

```
      Printout of the predicted survival from nursing home data

                                                                        Survivor
        Length                                Marriage Health Censoring Function
   Obs of stay age Treatment gender  status    index   index   Estimate

     1     1    69     0        0    Single     2       1      0.98969
     2     1    81     1        0    Single     2       1      0.98969
     3     1    93     0        0    Single     2       1      0.98969
     .     .     .     .        .       .       .       .       .
     .     .     .     .        .       .       .       .       .
     .     .     .     .        .       .       .       .       .
```

Calculating the median survival

Suppose that we want to calculate the median survival of single people with good health (i.e., `married=0` and `health=2`). There are 3 possible approaches to calculating the median:

1. Calculate the median from a specified covariate combination and perform a Kaplan-Meier analysis

   The problem with this approach is that we cannot do this for combinations where the covariate combination does not exist. For example, in the nursing home data, there are no individuals with health index 0 (i.e., totally healthy). So any combination in survival involving individuals with health index 0 cannot be performed by this approach.

2. Generate predicted survival curves for each combination of covariates and obtain the medians directly.

   For example, in the nursing home data we go to the printout above and look for the observations where the predicted survival (`predsurv`) goes from above 50% to below 50% in the specific group which we are interested in.

   For single persons with relatively good health (i.e., `health=2`), this is

Printout of the predicted survival from nursing home data

Obs	Length of stay	age	Treatment	gender	Marriage status	Health index	Censoring index	Survivor Function Estimate
163	182	75	0	1	Single	2	1	0.50295
164	189	87	1	0	Single	2	1	0.49812
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.

   So, for single healthy (`health=2`) individuals, the median survival is 189 days.

3. We may also generate the estimate from the model itself, by the formula

$$S(M;Z) = [S_0(M)]^{e^{-\beta Z_i}}$$

   so that the median M satisfies

$$S_0(M) = [0.5]^{e^{-\beta Z_i}}$$

   Suppose that we wanted to estimate the median for a single unhealthy (i.e., `health=5`) subject. Note that this value for `health` is not part of the data

set! Thus, the methods suggested in 1. and 2. above are not capable of generating this estimate. This estimate is

$$S_0(M) = [0.5]^{e^{-\beta Z_i}} = [0.5]^{e^{-(\beta_1 Z_1 + \beta_2 Z_2)}}$$

From the output of the PHREG procedure we know that β_2=0.165 and Z_1=0 (single person), so the above becomes

$$S_0(M) = [0.5]^{e^{-(\beta_1 Z_1 + \beta_2 Z_2)}} = [0.5]^{e^{-0.165(5)}} = 0.7380$$

This is the predicted survival distribution (i.e., the survival function associated with an individual with `married=0` and `health=5`)!!!

So we must find the predicted survival for a covariate combination that does not exist in the data. SAS provides these estimates directly (i.e., we do not need to estimate $S_0(t)$):

```
* Set covariate at single (married=0) and healthy (health=0);
data cov0;
     married=0;
     health=5;
run;
```

i.e., we've created a data set cov0 with `married=0` and `single=5`. Then we invoke PHREG as follows:

```
proc phreg data=nurshome;
     model los*fail(0)=married health;
     baseline out=outsurv survival=predsurv covariates=cov0/nomean;
     title 'PH regression analysis of nursing home data';
run;
```

Notice the code fragment

```
baseline out=outsurv survival=predsurv covariates=cov0/nomean;
```

By printing the data `outsurv` we get the predicted survival in the variable `predsurv` and if we look at the point where the predicted baseline survival goes from above 0.50 to below 0.50. We have

```
Printout of the predicted survival from nursing home data

   Obs    married    health    los    predsurv

     .        .          .        .          .
     .        .          .        .          .
     .        .          .        .          .
    80        0          5       80       0.50149
    81        0          5       81       0.49872
     .        .          .        .          .
     .        .          .        .          .
     .        .          .        .          .
```

so that the median is about 80 days.

The results of the printing procedure are as follows:

Printout of pred. survival from nursing home data
For single healthy persons

Obs	Marriage status	Health index	Length of stay	Survivor Function Estimate
1	0.17096	3.31615	0	1.00000
2	0.17096	3.31615	1	0.98653
3	0.17096	3.31615	2	0.97610
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
392	0.17096	3.31615	900	0.16566
393	0.17096	3.31615	911	0.16263
394	0.17096	3.31615	941	0.15920
395	0.17096	3.31615	942	0.15568

The question is why are there 395 observations in this data set? We note that

```
* Why are there 395 lines in the previous data set?
proc freq data=nurshome;
     table health*married;
         title 'Frequency table of married versus health status';
run;
```

Frequency table of married versus health status

The FREQ Procedure

Table of health by married

health(Health index)
married(Marriage status)

Frequency / Percent / Row Pct / Col Pct	Single	Married	Total
2	299	42	341
3	468	106	574
4	417	91	508
5	135	33	168
Total	1319	272	1591

So there are 299 failures in the single group with health index 2. So the question is why there are 395 entries in the table of predicted survival probabilities.

To answer this question we proceed with the following data code:

```
* Find the number of distinct LOS with at least one failure;
* STEP 1: Sort data by LOS;
proc sort data=prsurvsort out=lossort;
     by los;
run;

* STEP 2: find all LOS with AT LEAST one failure;
data lossortfirst;
     set lossort (keep=los fail);
     by los;                          *<-- Look within each group of LOS;
     retain found;
     if first.los then found=0;       *<-- At first LOS set found to zero;
     if fail=1 & found=0 then do;     *<-- Find the first LOS with a failure;
        output;                       *<-- Output it to the dataset;
        found=1;                      *<-- Set found to one (failure found);
     end;                             *<-- Only output the first failure;
     drop found fail;
run;                                  *<-- If no failures do not output;
```

We explain the code line by line below:

```
     set lossort (keep=los fail);
     by los;                          *<-- Look within each group of LOS;
```

By reading through the sorted data set `lossort` by `los`, we in effect read through groups of observations with the same `los` and operate on this group.

```
     retain found;
```

The command `retain`, copies `found` regardless of whether it has been updated at each line of the data set. In this way, we can create a variable which is updated on only a subset of the lines read from data set `lossort`. This is a very useful SAS command that you should keep in mind and use extensively! What we are trying to accomplish is have a code which tells SAS whether the first `los` in the group, with a failure, has been found since we will only want to output a single `los` from each group of length of stays which has at least one failure.

```
     if first.los then found=0;       *<-- At first LOS set found to zero;
```

At the first observation from a new `los` group we set `found` to zero (i.e., a failure has not been found).

```
     if fail=1 & found=0 then do;     *<-- Find the first LOS with a failure;
        output;                       *<-- Output it to the dataset;
        found=1;                      *<-- Set found to one (failure found);
     end;                             *<-- Only output the first failure;
```

The above code fragment says to SAS that we will output lines to the new data set only if `fail=1` (i.e., there is a failure at this observation) and if no failure has previously been found (i.e., `found=0`) in that `los` group. After outputting the observation to the new data set, `found` is set to one, which prevents any additional observations from the same `los` group from being output to the new

dataset. Also notice that, if no failure has been observed in the current `los` group, no observations will be output to the new data set. Thus, the data set should include only distinct `los` with at least one failure. Here is the log file summarizing the creation of data set `lossortfirst`.

```
NOTE: There were 1591 observations read from the data set WORK.LOSSORT.
NOTE: The data set WORK.LOSSORTFIRST has 394 observations and 1 variables.
NOTE: DATA statement used (Total process time):
      real time           0.01 seconds
      cpu time            0.01 seconds
```

So we see that the new data set `lossortfirst` has 394 observations. In other words, there are 394 distinct `los` with at least a single failure. The reason that the data set with the predicted probabilities has 395 observations is because SAS adds an observation with `los=0` and predicted survival estimate 1.0. To reassure ourselves that the new data set contains the same `los` entries as the one holding the predicted probabilities, we print the data below.

```
*STEP 3: Print the data set;
proc print data=lossortfirst;
     title 'Distinct LOS with at least one failure';
run;
```

Distinct LOS with at least one failure

Obs	los
1	1
2	2
3	3
.	.
.	.
.	.
392	911
393	941
394	942

So we see that the entries are identical to those listed under the data set of the predicted survival function (with the exception of course that there is no observation for `los=0`).