Session 6: Assessing the PH Assumption

In today’s lab, we are going to evaluate the assumption of proportional hazards using several graphical approaches. We will use the same example as in the lecture, the nursing home dataset (*nurshome.dat).* 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’;
     value sexfmt 0='Women' 1='Men';
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. Gender sexfmt.;
run;
```

We want to assess the proportional hazards for gender and marital status. One way is to produce the plot of *log[-log(S(t))]* versus *log(t).* If the lines of the subgroups are parallel then the assumption is satisfied. SAS produces this as follows:

```
proc phreg data=nurshome noprint;
     model los*fail(0)=gender;
     output out=outsurv loglogs=loglogsurv;
     title 'PH regression analysis of nursing home data';
run;
```

Note here the option `noprint` in the invocation of the command, which suppresses the output (this is frequently used when we are generating data sets with a procedure only). Also note that we now are outputting in the data set `outsurv` the log(-log(S(t)) by using the option `loglogs`.

To see what the data set looks like do the following:

```
proc print data=outsurv;run;
```

```
                   PH regression analysis of nursing home data

                    Obs    los    fail    gender    loglogsurv

                      1    665      1        0        0.37937
                      2    697      0        0        0.39442
                      3      7      1        0       -2.66415
                      4    217      1        0       -0.10862
                      .      .      .        .          .
                      .      .      .        .          .
                      .      .      .        .          .
```

To visualize the data we perform the following. First we generate a new data set with a variable `logt` to hold the logarithm of time (length of stay).

```
data plotdata;
     set outsurv;
     logt=log(los);
     label logt='Log of length of stay (days)';
run;
```

Then, we sort the data by log time in order to plot appropriately (this is rather critical).

```
proc sort data=plotdata;
     by logt;
run;
```

Then we perform the preparatory steps to generate the plot. Note that we simply join the points by using the following command:

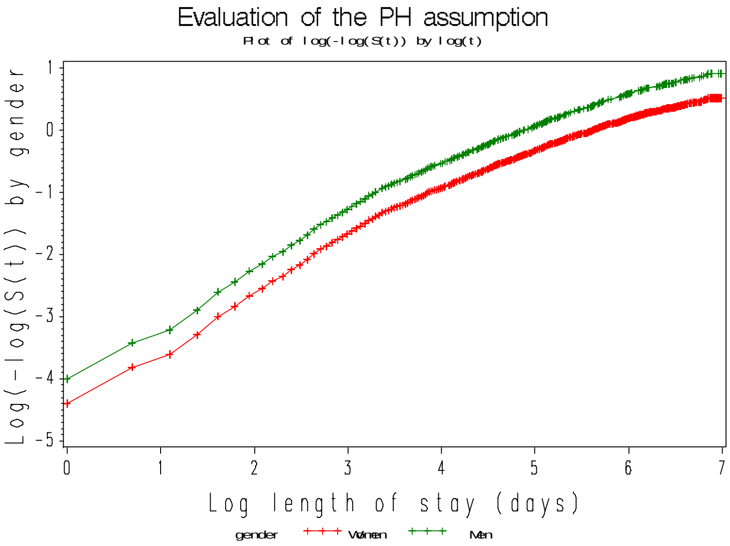
```
i=join
```

```
symbol1 c=red   line=1 i=join value=plus;
symbol2 c=green line=1 i=join value=plus;
```

```
axis1 label=(angle=90 height=2.0 font='arial'
             'Log(-log(S(t)) by gender' ) value=(font='arial'
      height=1.5);
axis2 label=(height=2.0 font='arial' 'Log length of stay (days)')
      value=(font='arial' height=1.5) minor=NONE;
```

```
proc gplot data=plotdata;
     plot loglogsurv*logt=gender/overlay vaxis=axis1 haxis=axis2;
     title 'Evaluation of the PH assumption';
     title2'Plot of log(-log(S(t)) by log(t)';
run;
```

The graph looks as follows:



Recall that SAS produces the log(-log(S(t)) (instead –log(-log(S(t)) as in STATA. This plot is equivalent to the STATA plot using the `noneg` option.

A similar graph can be created for Marital Status:

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

proc print data=outsurv;run;
```

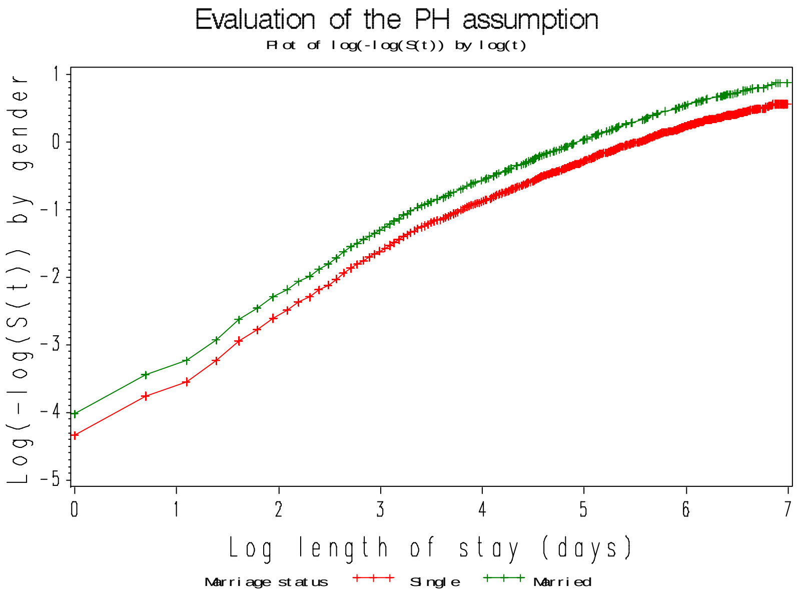
The data look as follows (notice that now you have marital status in the data set instead of gender).

```
                    PH regression analysis of nursing home data


                     Obs    los    fail    gender    loglogsurv


                       1    665      1     Women        0.37937
                       2    697      0     Women        0.39442
                       3      7      1     Women       -2.66415
                       4    217      1     Women       -0.10862
                       .     .       .       .           .
                       .     .       .       .           .
                       .     .       .       .           .
```

```
data plotdata;
     set outsurv;
     logt=log(los);
     label logt='Log of length of stay (days)';
run;

proc sort data=plotdata;
     by logt;
run;
```

```
symbol1 c=red   line=1 i=join value=plus;
symbol2 c=green line=1 i=join value=plus;

axis1 label=(angle=90 height=2.0 font='arial' 'Log(-log(S(t)) by
gender' )
      value=(font='arial' height=1.5);
axis2 label=(height=2.0 font='arial' 'Log length of stay (days)')
      value=(font='arial' height=1.5) minor=NONE;

proc gplot data=plotdata;
     plot loglogsurv*logt=married/overlay vaxis=axis1 haxis=axis2;
     title 'Evaluation of the PH assumption';
       title2'Plot of log(-log(S(t)) by log(t)';
       format gender sexfmt.;
run;
```

The plot is as follows:



Now we will generate Kaplan-Meier survival curves according to gender and compare them to the Cox predicted curves for the same variable. The closer the observed values are to the predicted, the less likely the proportional hazards assumption has been violated.

There are a number of steps that need to be performed in SAS in order to accomplish this (although not really complicated they are very time consuming compared especially to the STATA command `stcoxkm`).

We generate a new data set `outsurv` that, this time, contains the estimated survival data for the PH analysis by gender.

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

This will produce a new data set named `outsurv` with the Cox prediction of survival stored in variable `coxsurv`.

Now we need to run a Kaplan-Meier analysis and save the Kaplan-Meier prediction of the same survival data. We do this using PROC LIFETEST as we've done previously.

```
proc lifetest data=nurshome outsurv=sexkmsurv method=pl noprint;
     time los*fail(0);
       strata gender;
       title 'Kaplan Meier analysis of nursing home data';
run;
```

Notice the option `outsurv=sexkmsurv` in the invocation of the procedure. This produces a data set that holds the survival estimates of the Kaplan-Meier procedure and confidence intervals.

The data set is printed and looks as follows:

```
proc print data=sexkmsurv;run;
```

```
                       Kaplan Meier analysis of nursing home data

Obs    gender    los     _CENSOR_    SURVIVAL    SDF_LCL    SDF_UCL    STRATUM

   1     Men          0         0       1.00000    1.00000    1.00000       1
   2     Men          1         0       0.99282    0.98473    1.00000       1
   3     Men          2         0       0.97608    0.96143    0.99073       1
   4     Men          3         0       0.96172    0.94333    0.98012       1
   .       .          .         .          .          .          .          .
   .       .          .         .          .          .          .          .
   .       .          .         .          .          .          .          .
 238     Women        0         0       1.00000    1.00000    1.00000       2
 239     Women        1         0       0.98380    0.97658    0.99103       2
 240     Women        2         0       0.97528    0.96639    0.98416       2
   .       .          .         .          .          .          .          .
   .       .          .         .          .          .          .          .
   .       .          .         .          .          .          .          .
```

Unfortunately this is a different format compared to the previous data set (where each individual has an associated survival estimate.

We will need to create a common data set that is sorted by gender and then, within each gender, by the length of stay (`los`) variable and merge them together. This will be accomplished by a new data step where the data sets `sexkmsurv` and `outsurv` will be merged *by* `gender los`. But before we can do this we must sort the `sexoutsurv` data set by `los`; and not by increasing order of magnitude but by decreasing (as the Kaplan-Meier data set is thus sorted), while we also must sort the data by gender (in decreasing order as well, i.e., first men and then women). This is done as follows:

```
proc sort data=outsurv;
     by descending gender los;
run;

proc print data=outsurv; run;
```

The sorted data set looks like this:

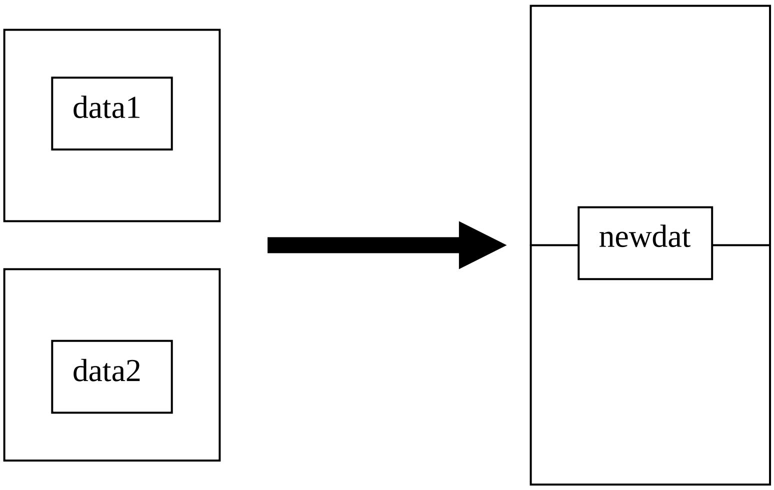
```
                      PH regression analysis of nursing home data

                        Obs     los    fail    gender    coxsurv

                          1       1       1       Men      0.98184
                          2       1       1       Men      0.98184
                          3       1       1       Men      0.98184
                          .       .       .        .        .
                          .       .       .        .        .
                          .       .       .        .        .
                        419       1       1     Women      0.98773
                        420       1       1     Women      0.98773
                        421       1       1     Women      0.98773
                          .       .       .        .        .
                          .       .       .        .        .
                          .       .       .        .        .
```

Now we are ready to merge the data. In SAS, there are two ways to merge two data sets: You can *set* them and you can *merge* them.

Using the set command, that is using code of the type

```
data newdata;
       set data1 data2 ;
       .
       .   *More SAS statements;
       .
run;
```

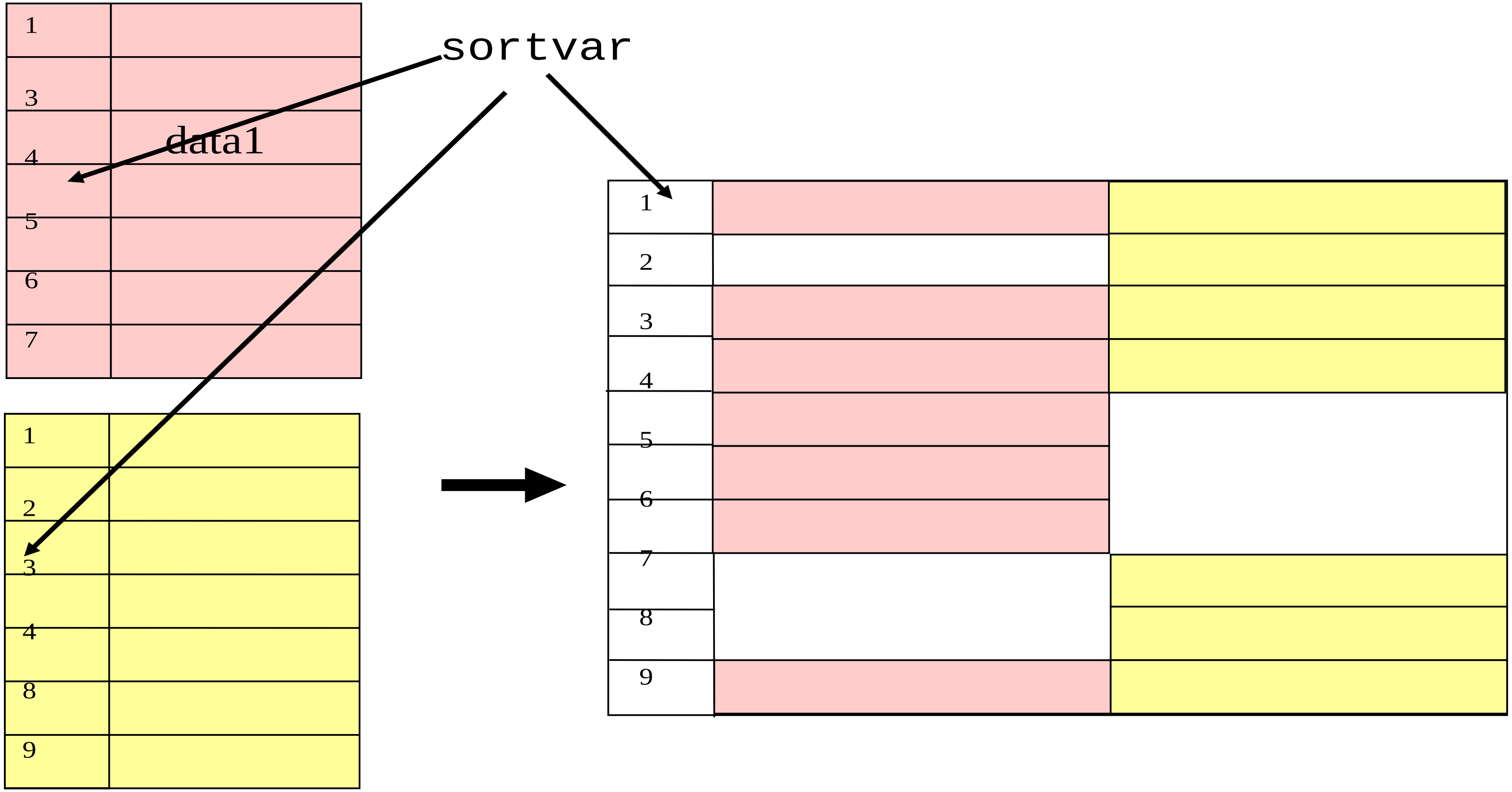
This accomplishes the following:



Merging two data sets on the other hand *by* a variable (say) `sortvar` is performed through the following SAS code (make sure that both data sets are sorted in the same order (ascending or descending) by sortvar:

```
data newdata;
       merge data1 data2 ;
       by sortvar;
       .
       .   *More SAS statements;
       .
run;
```

The results are given in the next page.



sortvar

data1

1
3
4
5
6
7

1
2
3
4
8
9

1
2
3
4
5
6
7
8
9

The SAS code to accomplish this is

```
data mergesurv;
     merge outsurv sexkmsurv;
       by descending gender los;
       if gender=1 then survival1=survival;
       if gender=0 then survival0=survival;
       if gender=1 then coxsurv1=coxsurv;
       if gender=0 then coxsurv0=coxsurv;
       label survival1='KM survival (males)'
             survival0='KM survival (females)'
               coxsurv1='Cox survival (males)'
               coxsurv0='Cox survival (females)';
       drop coxsurv survival _censor_ sdf_lcl sdf_ucl stratum;
run;

proc print data=mergesurv;
     title 'Merged survival data';
run;
```

The code above has the following results:

```
                              Merged survival data

  Obs     los     fail     gender     survival1     survival0     coxsurv1     coxsurv0
    1      0        .        Men        1.00000           .            .             .
    2      1        1        Men        0.99282           .         0.98184          .
    3      1        1        Men        0.99282           .         0.98184          .
    .      .        .         .           .               .            .             .
    .      .        .         .           .               .            .             .
    .      .        .         .           .               .            .             .
  420        0      .        Women        .             1.00000        .             .
  421        1      1        Women        .             0.98380        .          0.98773
  422        1      1        Women        .             0.98380        .          0.98773
     .       .      .         .           .               .            .             .
     .       .      .         .           .               .            .             .
     .       .      .         .           .               .            .             .
```

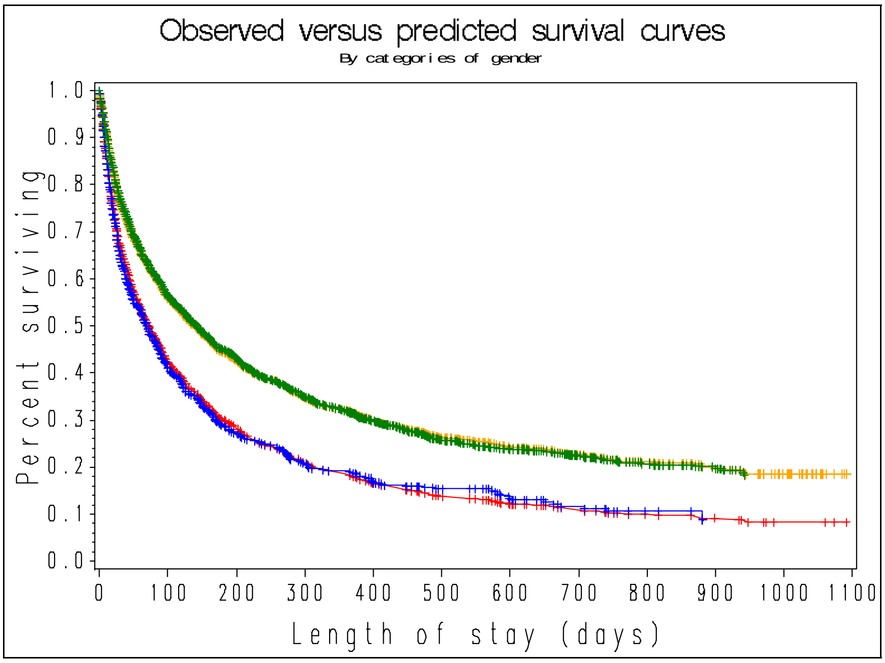
Now we can generate the Kaplan Meier and Cox plots and check the fit of the PH model. First we define the symbols and axis labels. Notice how we have decided to join the points in the first two symbols (that will eventually correspond to the Cox survival estimates (which are smooth), while we are producing a step function for the Kaplan-Meier curve.

```
symbol1 c=red    line=1 i=join     value=plus;
symbol2 c=orange line=1 i=join     value=plus;
symbol3 c=blue   line=1 i=stepljs  value=plus;
symbol4 c=green  line=1 i=stepljs  value=plus;
axis1 label=(angle=90 height=2.0 font='arial' 'Percent surviving' )
      value=(font='arial' height=1.5);
axis2 label=(height=2.0 font='arial' 'Length of stay (days)')
      value=(font='arial' height=1.5) minor=NONE;
```

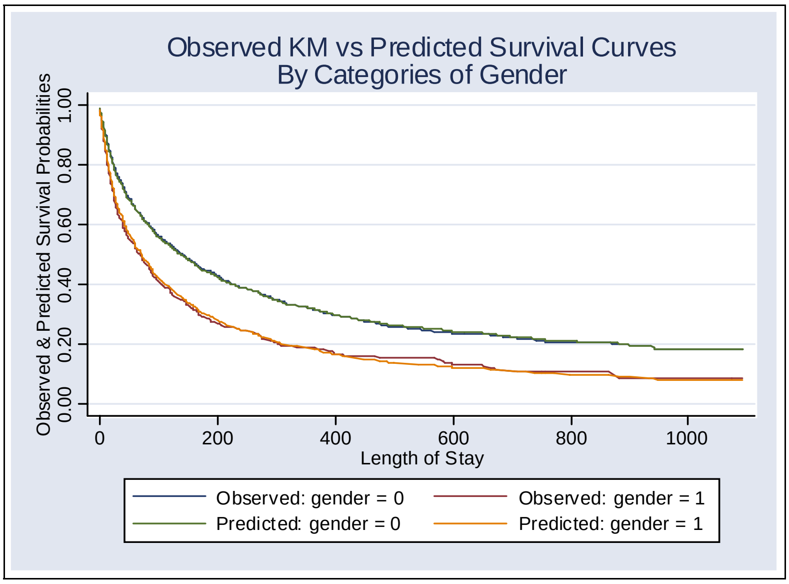
The GPLOT procedure in SAS is as follows:

```
proc gplot data=mergesurv;
     plot coxsurv1*los=1
          coxsurv0*los=2
          survival1*los=3
          survival0*los=4/overlay vaxis=axis1 haxis=axis2;
       title 'Observed versus predicted survival curves';
       title2 'By categories of gender';
run;
```

resulting in the following plot:



Compare this to the STATA plot


Observed KM vs Predicted Survival Curves
By Categories of Gender
Observed & Predicted Survival Probabilities
Length of Stay
Observed: gender = 0
Observed: gender = 1
Predicted: gender = 0
Predicted: gender = 1