Session 3: The PHREG procedure

We will first learn how to manipulate SAS data libraries. These are catalogs of data sets in the SAS format.

Consider the statement

```
libname datalib 'data';
```

This defines the library `datalib`, which is located in the subdirectory `\data` of the default directory.

In this library it is located the data set `mac`, which is already in SAS format. We will create a new data set `mac`, reading from this data set and only including a small subset of the variables in the original data set.

The data step statements are as follows:

```
data newmac;
     set datalib.mac (keep=patid macstat mactime rif clari cd4 karnof);
     label patid='Patient ID'
           macstat='Status of MAC infection'
           mactime='Time until MAC infection'
           rif='Rifabutin monotherapy'
           clari='Clarithromycin monotherapy'
           karnof='Karnofsky score'
           cd4='CD4+ count';
run;
```

Now consider the new statements. The statement that makes SAS read from a data set already created is the `set` statement, i.e.,

```
      set datalib.mac
```

Notice also, how the library information is conveyed. We put the library information before the data name in the library, followed by a period. Thus, the *new dataset* `newmac` will read from the SAS data set `mac` that is located in the library `datalib`.

We also do not want to read all the variables from the original dataset, so we include a `keep` statement along with the `set` command.

```
set datalib.mac (keep=patid macstat mactime rif clari cd4 karnof);
```

Alternatively, the keep statement could have been added as a command after the set statement as follows:

```
data newmac;
     set datalib.mac;
     keep patid macstat mactime rif clari cd4 karnof;
     label patid='Patient ID'
           macstat='Status of MAC infection'
           mactime='Time until MAC infection'
           rif='Rifabutin monotherapy'
           clari='Clarithromycin monotherapy'
           karnof='Karnofsky score'
           cd4='CD4+ count';
run;
```

Notice that there is no longer an equal sign ("=") or parentheses accompanying the keep statement. We will get the following comments in the log file

```
28   data newmac;
29        set datalib.mac (keep=patid macstat mactime rif clari cd4 karnof);
30        label patid='Patient ID'
31              macstat='Status of MAC infection'
32              mactime='Time until MAC infection'
33              rif='Rifabutin monotherapy'
34              clari='Clarithromycin monotherapy'
35              karnof='Karnofsky score'
36              cd4='CD4+ count';
37   run;

NOTE: There were 1177 observations read from the data set DATALIB.MAC.
NOTE: The data set WORK.NEWMAC has 1177 observations and 7 variables.
NOTE: DATA statement used:
      real time           0.03 seconds
      cpu time            0.03 seconds
```

To print the data set we write

```
options ls=80;
proc print data=newmac label;
     title 'The new mac data set';
run;
```

The output is as follows:

```
                            The new mac data set                              88
                                              07:27 Wednesday, December 10, 2003
                                 Status of Time until
      Patient Karnofsky  CD4+    MAC        MAC     Rifabutin  Clarithromycin
  Obs    ID     score   count infection  infection monotherapy   monotherapy

    1     1       90       8      1         560         1             0
    2     2       90      30      0         651         1             0
    3     3      100      80      0          26         1             0
    4     4       80      58      0         622         0             1
    5     5       90      59      0         643         0             1
    6     6       90      18      0         171         0             1
    7     7       90      20      1         174         0             0
    8     8       90      30      1         449         1             0
    9     9       80      30      1         377         0             0
   10    10       60      20      0          58         1             0
```

Notice the option statement that limits the width of the output to a line size of 80 columns

```
options ls=80;
```

Now let's carry out a proportional hazards regression with variables `rif`, `clari`, `karnof` and `cd4`. The SAS statements are as follows:

```
proc phreg data=newmac;
     model mactime*macstat(0)=rif clari karnof cd4;
     title 'PH regression analysis of the MAC data set';
run;
```

Notice the model statement of the PHREG procedure

```
     model mactime*macstat(0)=rif clari karnof cd4;
```

First comes the time variable, linked with the status (censoring/failure indicator) by an asterisk ("*"). Then the explanatory variables follow in the same manner as all regression procedures. The output is as follows:

```
                PH regression analysis of the MAC data set               105
                                      07:27 Wednesday, December 10, 2003

                          The PHREG Procedure

                           Model Information

     Data Set                  WORK.NEWMAC
     Dependent Variable        mactime          Time until MAC infection
     Censoring Variable        macstat          Status of MAC infection
     Censoring Value(s)        0
     Ties Handling             BRESLOW


           Summary of the Number of Event and Censored Values

                                                 Percent
                Total       Event    Censored    Censored

                 1177         121        1056       89.72

                          Convergence Status

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

                       Model Fit Statistics

                                  Without           With
                Criterion      Covariates     Covariates

                -2 LOG L         1541.064       1477.325
                AIC              1541.064       1485.325
                SBC              1541.064       1496.508
                Testing Global Null Hypothesis: BETA=0

          Test                 Chi-Square       DF     Pr > ChiSq

          Likelihood Ratio        63.7399        4         <.0001
          Score                   56.1915        4         <.0001
          Wald                    55.5623        4         <.0001

                  Analysis of Maximum Likelihood Estimates
```

Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq
rif	1	0.88034	0.23711	13.7846	0.0002
clari	1	0.25302	0.25835	0.9592	0.3274
karnof	1	-0.03685	0.01067	11.9405	0.0005
cd4	1	-0.01836	0.00368	24.8254	<.0001

Analysis of Maximum Likelihood Estimates

Variable	Hazard Ratio	Variable Label
rif	2.412	Rifabutin monotherapy
clari	1.288	Clarithromycin monotherapy
karnof	0.964	Karnofsky score
cd4	0.982	CD4+ count

First we obtain information about the convergence of the model, as well as information about the significance of the whole model (i.e., likelihood ratio tests, AIC, BIC factors that we use to compare between models and so on).

Wald tests (chi-square) and hazard ratios are given for all variables.