Self-Study Material 4: An example

The data that we are going to use is the UIC data

The goal of the UIS data is to model time until return to drug use for patients enrolled in two different residential treatment programs that differed in length (treat=0 is the short program and treat=1 is the long program). The patients were randomly assigned to two different sites (site=0 is site A and site=1 is site B). The variable age indicates age at enrollment, herco indicates heroine or cocaine use in the past three months (herco=1 indicates heroine and cocaine use, herco=2 indicates either heroine or cocaine use and herco=3 indicates neither heroine nor cocaine use) and ndrugtx indicates the number of previous drug treatments. The variables time contains the time until return to drug use and the censor variable indicates whether the subject returned to drug use (censor=1 indicates return to drug use and censor=0 otherwise).

Note that the coding for censor is rather counter-intuitive since the value 1 indicates an event and 0 indicates censoring. It would perhaps be more appropriate to call this variable "event".

2.1 Exploring the data: Univariate Analyses

In any data analysis it is always a great idea to do some univariate analysis before proceeding to more complicated models. In survival analysis it is highly recommended to look at the Kaplan-Meier curves for all the categorical predictors. This will provide insight into the shape of the survival function for each group and give an idea of whether or not the groups are proportional (i.e. the survival functions are approximately parallel). We also consider the tests of equality across strata to explore whether or not to include the predictor in the final model. For the categorical variables we will use the log-rank test of equality across strata, which is a non-parametric test. For the continuous variables we will use a univariate Cox proportional hazard regression which is a semi-parametric model. We will consider including the predictor if the test has a p-value of 0.2 - 0.25 or less. We are using this elimination scheme because all the predictors in the data set are variables that could be relevant to the model. If the predictor has a p-value greater than 0.25 in a univariate analysis it is highly unlikely that it will contribute anything to a model which includes other predictors.

The log-rank test of equality across strata for the predictor treat has a p-value of 0.0091. From the graph we see that the survival function for each group of treat are not perfectly parallel but separate except at the very beginning and at the very end.

stset time, failure(censor==1)

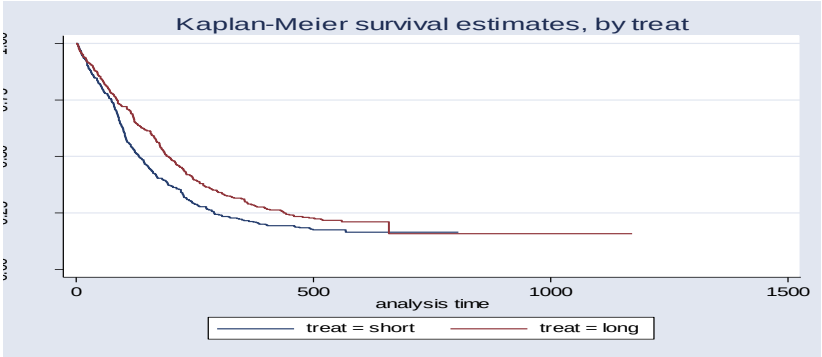
```
sts test treat, logrank

Log-rank test for equality of survivor functions

      |   Events         Events
```

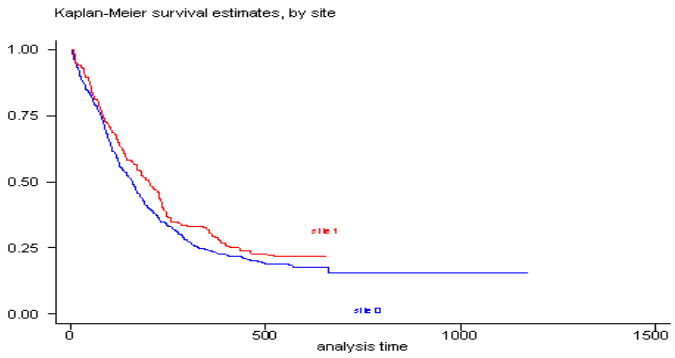
```
treat |  observed       expected
------+-------------------------
0     |       265         235.80
1     |       243         272.20
------+-------------------------
Total |       508         508.00

           chi2(1) =       6.80
           Pr>chi2 =     0.0091
sts graph, by(treat)
```



The log-rank test of equality across strata for the predictor site has a p-value of 0.1240, thus site will be included as a potential candidate for the final model because this p-value is still less than our cut-off of 0.2. From the graph we see that the survival curves are not all that parallel and that there are two periods ([0, 100] and [200, 300]) where the curves are very close together. This would explain the rather high p-value from the log-rank test.

```
sts test site, logrank
sts graph, by(site)
         failure _d:  censor
   analysis time _t:  time
Log-rank test for equality of survivor functions
      |   Events         Events
site  |  observed       expected
------+-------------------------
0     |       364         347.94
1     |       144         160.06
------+-------------------------
Total |       508         508.00
           chi2(1) =       2.37
           Pr>chi2 =     0.1240
```



The log-rank test of equality across strata for the predictor herco has a p-value of 0.1473, thus herco will be included as potential candidate for the final model. From the graph we see that the three groups are not parallel and that especially the groups herco=1 and herco=3 overlap for most of the graph. This lack of parallelism could pose a problem when we include this predictor in the Cox proportional hazard model since one of the assumptions is proportionality of the predictors.

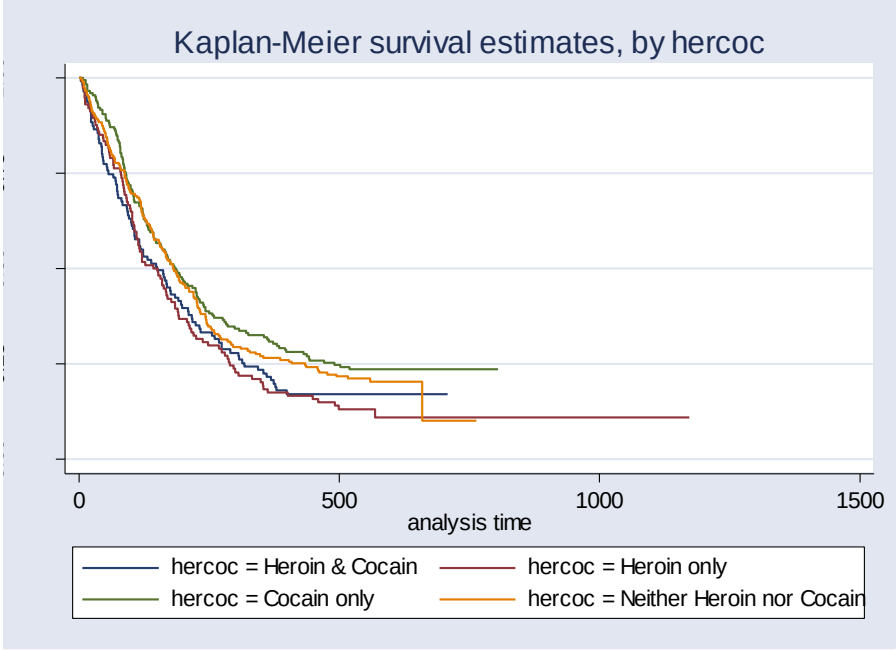
```
sts graph, by(herco)
 . sts test herco

         failure _d:  censor == 1
   analysis time _t:  time

Log-rank test for equality of survivor functions

                          |    Events         Events
hercoc                    |  observed       expected
--------------------------+-------------------------
Heroin & Cocain           |        92          81.91
Heroin only               |       100          82.43
Cocain only               |       136         156.05
Neither Heroin nor Cocain |       165         172.61
--------------------------+-------------------------
Total                     |       493         493.00

                                chi2(3) =       7.95
                                Pr>chi2 =     0.0470
```



It is not feasible to calculate a Kaplan-Meier curve for the continuous predictors since there would be a curve for each level of the predictor and a continuous predictor simply has too many different levels. Instead we consider the Cox proportional hazard model with a single continuous predictor. Unfortunately it is not possibly to produce a plot when using the stcox command. Instead we consider the Chi-squared test for ndrugtx which has a p-value of 0.0003 thus ndrugtx is a potential candidate for the final model since the p-value is less than our cut-off value of 0.2. We specify the option nohr to indicate that we do not want to see the hazard ratio rather we want to look at the coefficients.

```
stcox ndrugtx, nohr

Cox regression -- Breslow method for ties
No. of subjects =          611                     Number of obs    =       611
No. of failures =          496
Time at risk    =       143002
                                                   LR chi2(1)       =     13.35
Log likelihood  =    -2868.299                     Prob > chi2      =    0.0003
------------------------------------------------------------------------------
          _t |
          _d |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     ndrugtx |    .029372   .0074979     3.92   0.000     .0146763    .0440676
------------------------------------------------------------------------------
```

In this model the Chi-squared test of age also has a p-value of less than 0.2 and so it is a potential candidate for the final model.

```
stcox age, nohr
```

```
Cox regression -- Breslow method for ties
------------------------------------------------------------------------------
          _t |
          _d |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.0128641   .0071888    -1.79   0.074    -.0269539    .0012256
------------------------------------------------------------------------------
```

2.3 Model Building

For our model building, we will first consider the model which will include all the predictors that had a p-value of less than 0.2 - 0.25 in the univariate analyses which in this particular analysis means that we will include every predictor in our model. The categorical predictor herco has three levels and therefore we will include this predictor using dummy variable with the group herco=1 as the reference group. We can create these dummy variables on the fly by using the xi command with coxreg.

```
. xi: stcox age ndrugtx treat site i.herco, nohr
i.hercoc          _Ihercoc_1-4        (naturally coded; _Ihercoc_1 omitted)

         failure _d:  censor == 1
   analysis time _t:  time

Iteration 0:   log likelihood =   -2773.97
Iteration 1:   log likelihood = -2755.1644
Iteration 2:   log likelihood = -2754.5507
Iteration 3:   log likelihood = -2754.5486
Refining estimates:
Iteration 0:   log likelihood = -2754.5486

Cox regression -- Breslow method for ties

No. of subjects =          593                     Number of obs   =       593
No. of failures =          481
Time at risk    =       141069
                                                   LR chi2(7)      =     38.84
Log likelihood  =   -2754.5486                     Prob > chi2     =    0.0000

------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.0279101   .0077907    -3.58   0.000    -.0431795   -.0126407
     ndrugtx |   .0346947   .0078855     4.40   0.000     .0192393      .05015
       treat |  -.2502843   .0923905    -2.71   0.007    -.4313665   -.0692022
```

```
        site |  -.0984204   .1038916    -0.95   0.343    -.3020442    .1052034
  _Ihercoc_2 |   .1140884   .1464417     0.78   0.436     -.172932    .4011088
  _Ihercoc_3 |  -.2404967   .1416395    -1.70   0.090     -.518105    .0371117
  _Ihercoc_4 |  -.0884732   .1383712    -0.64   0.523    -.3596757    .1827293
------------------------------------------------------------------------------
.  test _Ihercoc_2 _Ihercoc_3 _Ihercoc_4

 ( 1)  _Ihercoc_2 = 0
 ( 2)  _Ihercoc_3 = 0
 ( 3)  _Ihercoc_4 = 0

           chi2(  3) =    7.07
         Prob > chi2 =    0.0697
```

The predictors herco and site are not significant and we will drop them from the final model. So, the final model of main effects include: age, ndrugtx and treat.

```
. stcox age ndrugtx treat, nohr

         failure _d:  censor == 1
   analysis time _t:  time
Cox regression -- Breslow method for ties

No. of subjects =          610                     Number of obs   =       610
No. of failures =          495
Time at risk    =       142994
                                                   LR chi2(3)      =     27.76
Log likelihood  =   -2854.6735                     Prob > chi2     =    0.0000

------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.0207666   .0074199    -2.80   0.005    -.0353094   -.0062238
     ndrugtx |   .0354906   .0076196     4.66   0.000     .0205564    .0504247
       treat |   -.230559   .0901757    -2.56   0.011    -.4073001   -.0538179
------------------------------------------------------------------------------
```

Next we need to consider interactions. We do not have any prior knowledge of specific interactions that we must include so we will consider all the possible interactions. Since our model is rather small this is manageable but the ideal situation is when all models building, including interactions, are theory driven.

```
gen age_drug = age*ndrugtx
gen age_treat = age*treat
gen treat_drug=treat*ndrugtx
```

```
None of these interactions is significant, so the final model does not include any
interaction.
------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .9794475   .0072674    -2.80   0.005     .9653067    .9937955
     ndrugtx |   1.036128   .0078949     4.66   0.000     1.020769    1.051718
       treat |   .7940896   .0716076    -2.56   0.011     .6654445    .9476047
------------------------------------------------------------------------------
```

From looking at the hazard ratios (also called relative risks) the model indicates that as the number of previous drug treatment (ndrugtx) increases by one unit, and all other variables are held constant, the rate of relapse increases by 3.6%. If the treatment length is altered from short to long, while holding all other variables constant, the rate of relapse decreases by (100% - 79.4%) = 20.6%. If age is increased by 10 years and all other variables are held constant the hazard ratio is equal to exp(-0.02*10) = .81. Thus, the rate of relapse is decreased by (100% - 81%) = 19% with an increase of 10 years in age.

2.3 Proportionality Assumption

One of the main assumptions of the Cox proportional hazard model is proportionality. There are several methods for verifying that a model satisfies the assumption of proportionality. We will check proportionality by using the Schoenfeld and scaled Schoenfeld residuals which must first be saved through the coxreg command. In the stphtest command we test the proportionality of the model as a whole and by using the detail option we get a test of proportionality for each predictor. By using the plot option we can also obtain a graph of the scaled Schoenfeld assumption. **If the tests in the table are not significance (p-values over 0.05) then we can not reject proportionality** and we assume that we do not have a violation of the proportional assumption. A horizontal line in the graphs is further indication that there is no violation of the proportionality assumption. The stphplot command uses log-log plots to test proportionality and if the lines in these plots are parallel then we have further indication that the predictors do not violate the proportionality assumption.

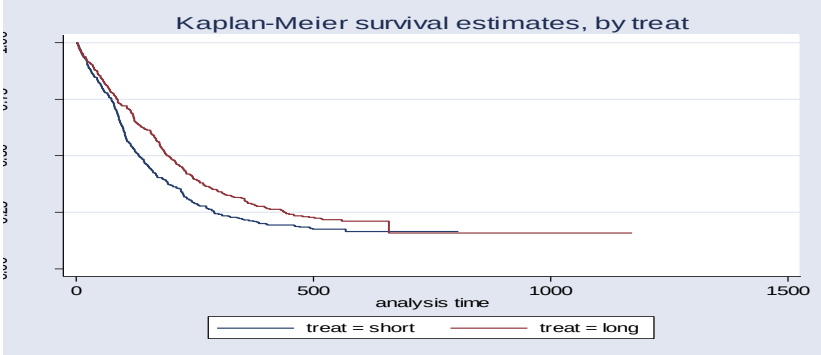
```
quietly stcox age ndrugtx treat, schoenfeld(sch*) scaledsch(sca*)
stphtest, det

      Test of proportional hazards assumption

      Time:  Time
      ----------------------------------------------------------------
                  |       rho            chi2       df       Prob>chi2
      ------------+---------------------------------------------------
      age         |      0.00507         0.01        1         0.9133
      ndrugtx     |      0.05127         1.23        1         0.2680
      treat       |      0.10432         5.34        1         0.0209
```

```
------------+------------------------------------------------------------
global test |                        6.86       3         0.0765
------------------------------------------------------------------------
```

The predictor treat might warrant some closer examination since it does have a significant test.



Although the two curves are not completely parallel, they are almost parallel except at the very beginning and at the very end. Also the graph doesn't have any cross-over. So we choose to leave treat in the model unaltered based on prior research.

If one of the predictors were not proportional there are various solutions to consider. One solution is to include the time-dependent variable for the non-proportional predictors. Another solution is to stratify on the non-proportional predictor. The following is an example of stratification on the predictor treat. Note that treat is no longer included in the model statement instead it is specified in the strata statement.

```
sort treat
by treat: stcox age ndrugtx, nohr
```

```
-------------------------------------------------------------------------------
-> treat = short
------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.0098603   .0102704    -0.96   0.337    -.0299899    .0102692
     ndrugtx |   .0365166   .0112408     3.25   0.001     .0144849    .0585482
------------------------------------------------------------------------------


-------------------------------------------------------------------------------
-> treat = long

------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.0344729   .0108633    -3.17   0.002    -.0557646   -.0131813
     ndrugtx |   .0361074    .010496     3.44   0.001     .0155355    .0566792
```

In the stratification model, one will obtain separate baseline hazard functions for each value of the categorical variable. One would do this, of course, if one thought that different categories had different baseline functions which were not proportional (if they were proportional, one could use the would-be stratification variable as a covariate; proportionality may be checked by Log-Minus-Log survival plots). The stratification variable is not treated as a predictor and no coefficients are computed for it.