Applied Survival Analysis
Lab 6: Model Selection in Survival Analysis

Today, we are going to understand Collet's approach for model selection within the context of a proportional hazards model and to assess the overall fit of the model by checking the residuals.

1. ***Collet's Approach for Model Selection:***

We are going to work with the MAC dataset (*mac.dta*), focusing on the outcome **`dthstat`** which equals 1 if a patient died and 0 otherwise. The time to death is **`dthtime`** and subjects who did not die are censored at their time of study discontinuation.
The covariates of interest for the purpose of this lab are:

agecat	**sex**	**cd4**	**karnof**
ivdrug	**antiret**	**rif**	**clari**

The significance of their effect will be tested using the Wald test. This time we are interested in the time to death. So we are going to *stset* the data in the following way:

```
stset  dthtime, failure(dthstat)
```

Step 1: Fit univariate models to choose candidate predictors. Use criterion of $p \le 0.15$ to identify predictors.

```
stcox agecat, nohr
```

```
         failure _d:  dthstat
   analysis time _t:  dthtime

Iteration 0:   log likelihood = -3393.2516
Iteration 1:   log likelihood = -3385.2229
Iteration 2:   log likelihood = -3385.2133
Refining estimates:
Iteration 0:   log likelihood = -3385.2133

Cox regression -- Breslow method for ties

No. of subjects =        1175                     Number of obs   =      1175
No. of failures =         514
Time at risk    =      619081
                                                  LR chi2(1)      =     16.08
Log likelihood  =   -3385.2133                    Prob > chi2     =    0.0001

------------------------------------------------------------------------------
      _t |
      _d |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
  agecat |   .3663803   .0928965      3.944   0.000       .1843065    .5484541
```

```
stcox sex, nohr

         failure _d:  dthstat
   analysis time _t:  dthtime

Iteration 0:   log likelihood = -3393.2516
Iteration 1:   log likelihood = -3392.0249
Iteration 2:   log likelihood = -3392.0109
Iteration 3:   log likelihood = -3392.0109
Refining estimates:
Iteration 0:   log likelihood = -3392.0109

Cox regression -- Breslow method for ties

No. of subjects =         1175                     Number of obs   =      1175
No. of failures =          514
Time at risk    =       619081
                                                   LR chi2(1)      =      2.48
Log likelihood  =   -3392.0109                     Prob > chi2     =    0.1152

------------------------------------------------------------------------------
      _t |
      _d |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
     sex |   .2360791   .1453047      1.625   0.104       -.048713    .5208711
```

(a) Fit all other univariate models of interest and fill in the below summary table of univariate predictors:

Predictor	***Estimate***	***s.e.***	***p-value***	***HR***
agecat				
sex				
cd4				
karnof				
ivdrug				
antiret				
rif				
clari				

Is the effect of therapy with rif and/or clari significant?

<u>Step 2:</u>

(i) Fit a multivariate model with all significant predictors ($p \leq 0.15$) from Step 1:

(ii) Then use backwards selection to eliminate non-significant ones in a multivariate framework (use $p \leq 0.10$ for determining which ones to eliminate).
(To use automatic variable selection in STATA, we use the **`sw stcox`** command).

```
sw stcox agecat sex cd4 karnof antiret, pr(0.10)
(2 obs. dropped due to estimability)
    LR test           begin with full model
p < 0.1000            for all terms in model

Cox regression -- Breslow method for ties

No. of subjects =         1175                     Number of obs   =      1175
No. of failures =          514
Time at risk    =       619081
                                                   LR chi2(5)      =    149.78
Log likelihood  =   -3318.3627                     Prob > chi2     =    0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      agecat |   1.420866   .1334364     3.74   0.000     1.181993    1.708013
         sex |   1.369403   .2001345     2.15   0.031     1.028326    1.823611
         cd4 |   .9895408    .001542    -6.75   0.000     .9865232    .9925676
      karnof |   .9626493   .0049101    -7.46   0.000     .9530736    .9723211
     antiret |   .7926794   .0786474    -2.34   0.019      .652595    .9628339
------------------------------------------------------------------------------
```

Step 3:

Use forwards selection to add any variables not significant at Step 1 to the multivariate model obtained at the end of Step 2. To be conservative, use $p \leq 0.10$ for deciding whether to add variables or not.

Note that we specify the option **`lockterm1`** to force the variables in the first parenthesis into the model.

```
sw stcox (agecat sex cd4 karnof antiret) ivdrug (rif clari), pe(0.10)
lockterm1
(2 obs. dropped due to estimability)
                      begin with term 1 model
p >=0.1000            for all terms in model

Cox regression -- Breslow method for ties

No. of subjects =         1175                     Number of obs   =      1175
No. of failures =          514
Time at risk    =       619081
                                                   LR chi2(5)      =    149.78
Log likelihood  =   -3318.3627                     Prob > chi2     =    0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      agecat |   1.420866   .1334364     3.74   0.000     1.181993    1.708013
         sex |   1.369403   .2001345     2.15   0.031     1.028326    1.823611
         cd4 |   .9895408    .001542    -6.75   0.000     .9865232    .9925676
      karnof |   .9626493   .0049101    -7.46   0.000     .9530736    .9723211
     antiret |   .7926794   .0786474    -2.34   0.019      .652595    .9628339
------------------------------------------------------------------------------
```

(b) Are there any other variables added to the model?

Step 4:

Create all possible 2-way interaction terms based on the main effects in the model at the end of Step 3. Add these to the multivariate model and use a backwards selection procedure to eliminate those not significant at $p = 0.10$. Remember to force the inclusion of all of the main effects from the end of Step 3 in the model.

First we are going to generate all possible 2-way interactions:

gen agsex=agecat*sex

gen agcd4=agecat*cd4

gen agkar=agecat*karnof

gen aganti=agecat*antiret

gen sexcd4=sex*cd4

gen sexkar=sex*karnof

gen sexanti=sex*antiret

gen cd4kar=cd4*karnof

gen cd4anti=cd4*antiret

gen karanti=karnof*antiret

(c) Now fit the appropriate model that was described above in Step 4.

<u>Step 5:</u>
Check if all variables are significant. If a main effect has become non-significant and there are no interactions involving this main effect in the model at the end of Step 4, then you may consider excluding it. Discretion is needed in determining whether to include covariates and/or interactions that are marginally significant. At this stage, for the purposes of this exercise, use a somewhat stricter significance level of $\alpha = 0.02$ to account for the multiple tests we have conducted.

sw stcox agecat sex cd4 karnof antiret sexanti karanti , pe(0.02)

```
(2 obs. dropped due to estimability)
                      begin with empty model
p = 0.0000 <  0.0200  adding   karnof
p = 0.0000 <  0.0200  adding   cd4
p = 0.0002 <  0.0200  adding   agecat

Cox regression -- Breslow method for ties

No. of subjects =         1175                     Number of obs   =      1175
No. of failures =          514
Time at risk    =       619081
                                                   LR chi2(3)      =    140.91
Log likelihood  =   -3322.7981                     Prob > chi2     =    0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      karnof |   .9628398   .0049084    -7.43   0.000     .9532673    .9725083
         cd4 |   .9895551   .0015446    -6.73   0.000     .9865324    .9925871
      agecat |   1.420961   .1330381     3.75   0.000     1.182736    1.707168
------------------------------------------------------------------------------
```

Step 6:

You may also want to include some models with and without certain marginally significant covariates to evaluate the change in the AIC criterion or consider alternate codings of covariates (eg. cd4cat instead of cd4 or age instead of agecat).

This time we could use the stepwise backwards procedure with probability of entry *pe* = 0.05 and probability of removal *pr* = 0.10.

(i) (cd4cat instead of cd4)

sw stcox agecat sex cd4cat karnof antiret, pe(0.05) pr(0.10)

```
(2 obs. dropped due to estimability)
                      begin with full model
p < 0.1000            for all terms in model

Cox regression -- Breslow method for ties

No. of subjects =         1175                     Number of obs   =      1175
No. of failures =          514
Time at risk    =       619081
                                                   LR chi2(5)      =    132.52
Log likelihood  =   -3326.9922                     Prob > chi2     =    0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      agecat |   1.414087   .1329843     3.68   0.000     1.176053    1.700299
         sex |   1.390386   .2033177     2.25   0.024      1.04391    1.851857
      cd4cat |   .5762558   .0525747    -6.04   0.000     .4818989     .689088
      karnof |   .9627031   .0049187    -7.44   0.000     .9531107     .972392
     antiret |   .7910011   .0785034    -2.36   0.018      .651177    .9608491
------------------------------------------------------------------------------
```

(ii) (age instead of agecat)

In this case we are including age instead of agecat. To use the stepwise forwards procedure we have to add the option **`forward`** (the default is backwards):

sw stcox age sex cd4 karnof antiret, pe(0.05) pr(0.10) forward

```
                      begin with empty model
p = 0.0000 <  0.0500  adding   karnof
p = 0.0000 <  0.0500  adding   cd4
p = 0.0000 <  0.0500  adding   age
p = 0.0406 <  0.0500  adding   sex
p = 0.0351 <  0.0500  adding   antiret

Cox regression -- Breslow method for ties
Entry time 0                                      Number of obs   =       1177
                                                  LR chi2(5)      =     153.07
                                                  Prob > chi2     =     0.0000
Log likelihood = -3316.7163                       Pseudo R2       =     0.0226

------------------------------------------------------------------------------
 dthtime |
 dthstat |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
  karnof | -.0376065   .0051224     -7.342   0.000      -.0476463   -.0275666
     cd4 | -.0105009   .0015625     -6.720   0.000      -.0135635   -.0074384
     age |  .0211124   .0049501      4.265   0.000       .0114104    .0308145
     sex |  .3210042   .1460386      2.198   0.028       .0347737    .6072346
 antiret | -.2099265   .0996185     -2.107   0.035      -.4051752   -.0146778
------------------------------------------------------------------------------
```

(d) Summarize the final models at the ends of Steps 1-6 in a table, such as that shown below. Use $\alpha = 3$ in calculating the AIC value ($AIC = -2logL + (\alpha * q)$) for each of the models below. Which model appears best in terms of the AIC criterion?

Model	***Covariates***	***-2logL***	***q***	***AIC***
Step 2 (i)	agecat, karnof, sex, cd4, antiret	6636.7	5	
Step 2 (ii)	(same as above)			
Step 3	(same as above)			
Step 4	Step3 +karnof*antiret,sex*antiret	6628.6	7	
Step 5	agecat, karnof, cd4	6645.6	3	
Step 6 (i)	agecat, karnof, sex, cd4cat, antiret	6654.0	5	
Step 6 (ii)	age, karnof, sex, cd4, antiret	6633.4	5	

2. ***<u>Assessing Overall Model Fit:</u>***

We will assess the overall fit of the model in Step 6 (ii) .

- **<u>Martingale Residuals:</u>**

First we will fit the model along with the option **`mgale (newvar)`** to get the *Martingale* residuals:

`stcox  age sex cd4 karnof antiret , mgale(mg)`

```
         failure _d:  dthstat
   analysis time _t:  dthtime

Iteration 0:   log likelihood = -3393.2516
Iteration 1:   log likelihood = -3318.6397
Iteration 2:   log likelihood = -3316.7201
Iteration 3:   log likelihood = -3316.7163
Refining estimates:
Iteration 0:   log likelihood = -3316.7163

Cox regression -- Breslow method for ties

No. of subjects =         1175                     Number of obs   =      1175
No. of failures =          514
Time at risk    =       619081
                                                   LR chi2(5)      =    153.07
Log likelihood  =   -3316.7163                     Prob > chi2     =    0.0000

------------------------------------------------------------------------------
      _t |
      _d | Haz. Ratio   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
     age |   1.021337   .0050557      4.265   0.000       1.011476    1.031294
     sex |   1.378511   .2013159      2.198   0.028       1.035385    1.835349
     cd4 |    .989554   .0015462     -6.720   0.000       .9865281     .9925892
  karnof |   .9630919   .0049334     -7.342   0.000        .953471     .9728098
 antiret |   .8106438   .0807551     -2.107   0.035         .66686     .9854294
```

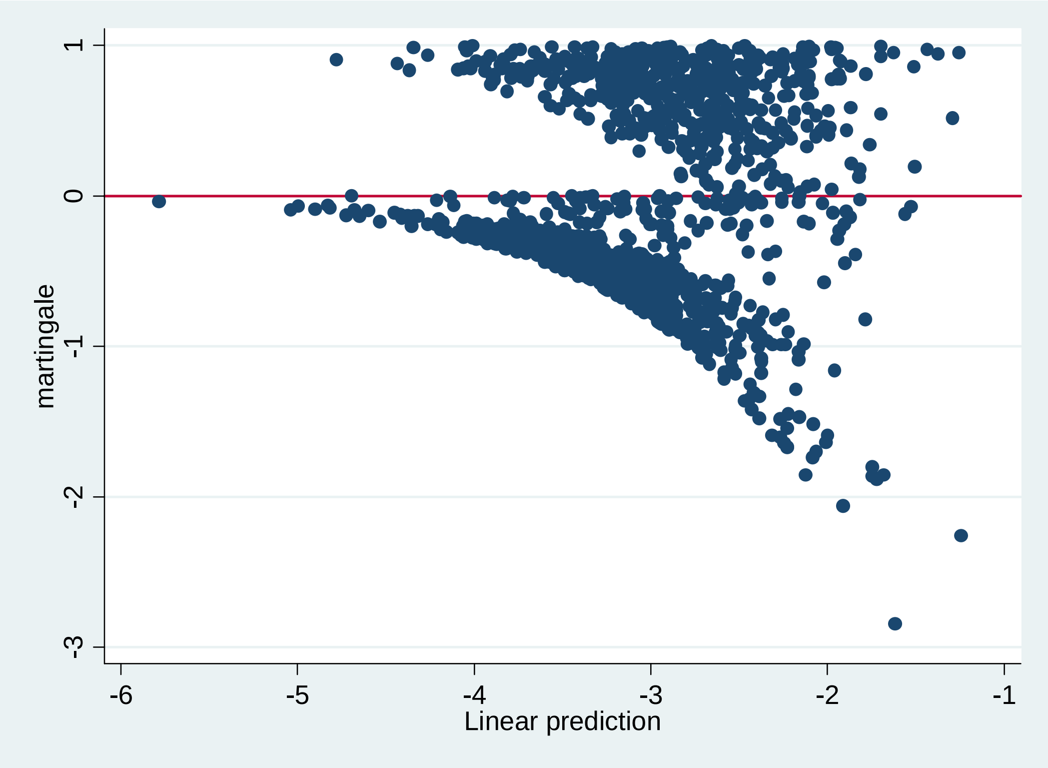
Once the martingale residuals are created, you usually plot them versus the predicted log HR or any of the individual covariates to assess the model fit.

To get the predicted log HR we use the following command:

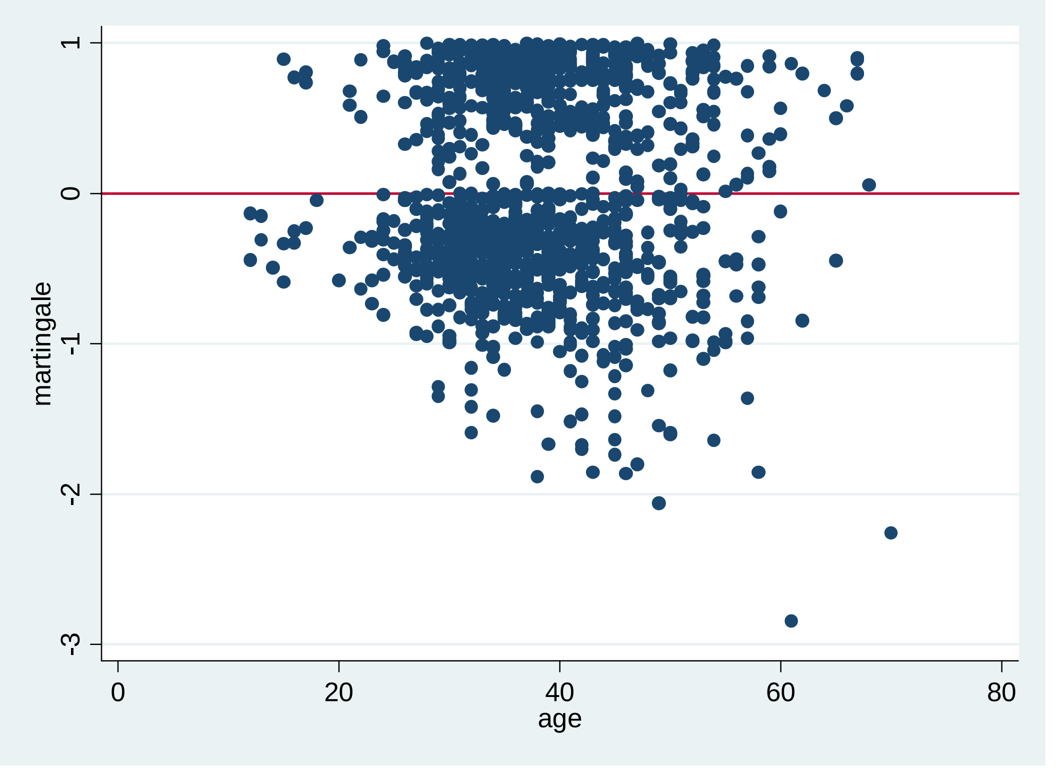
```
predict betaz, xb
 (2 missing values generated)
```

And the graphs:

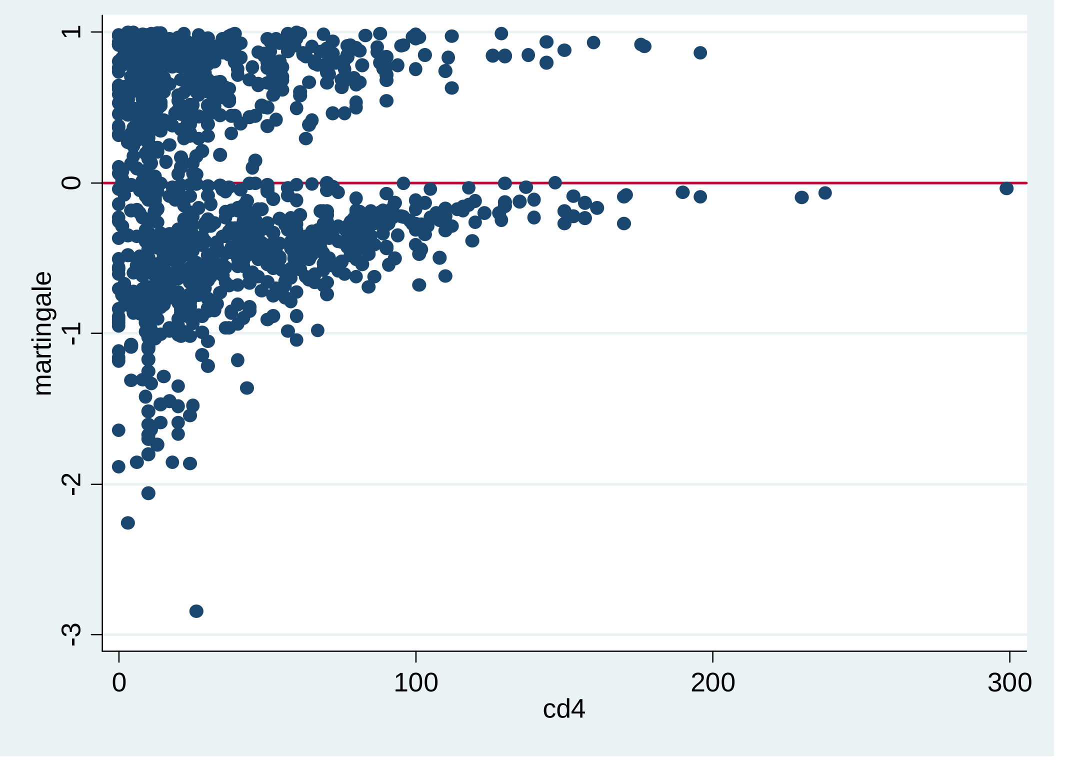
```
scatter mg betaz, yline(0)
```



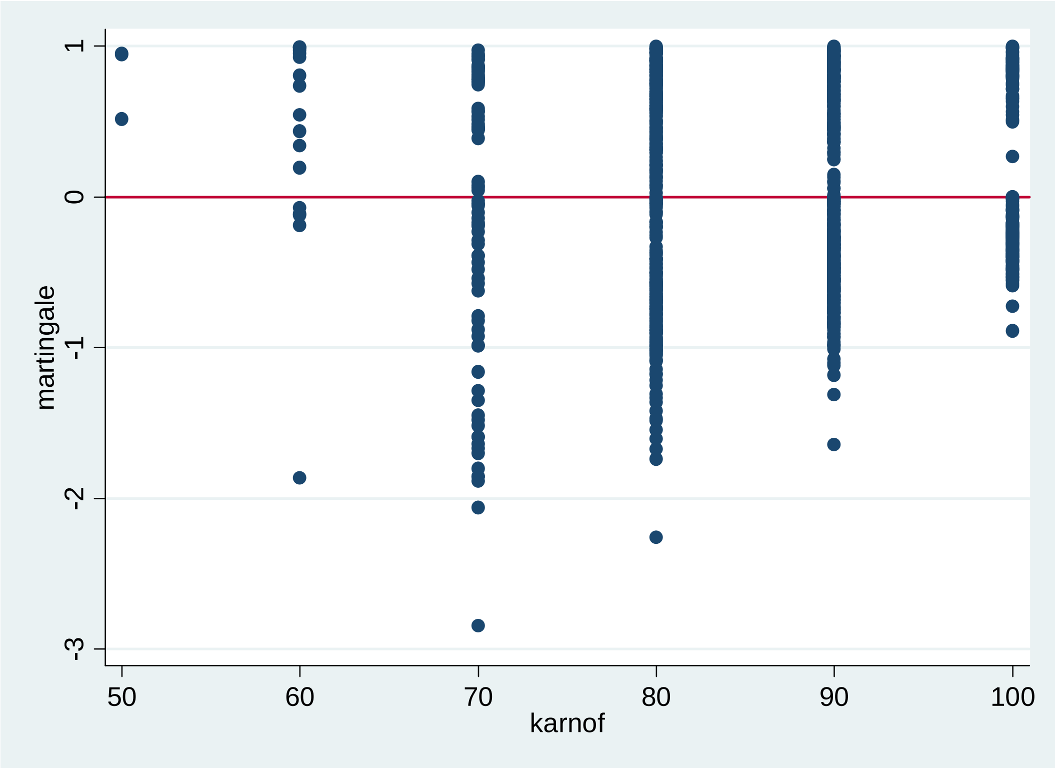
```
scatter mg age, yline(0)
```




```
scatter mg cd4, yline(0)
```



```
scatter mg karnof, yline(0)
```



♦ **Generalized (Cox-Snell) Residuals:**

To get generalized residuals in STATA we use the previous **`stcox`** command with the **`mgale`** option and then use the **`predict`** command with **`csnell`** option:

```
stcox  age sex cd4 karnof antiret , mgale(mg)
```

```
predict csres, csnell
(2 missing values generated)
```

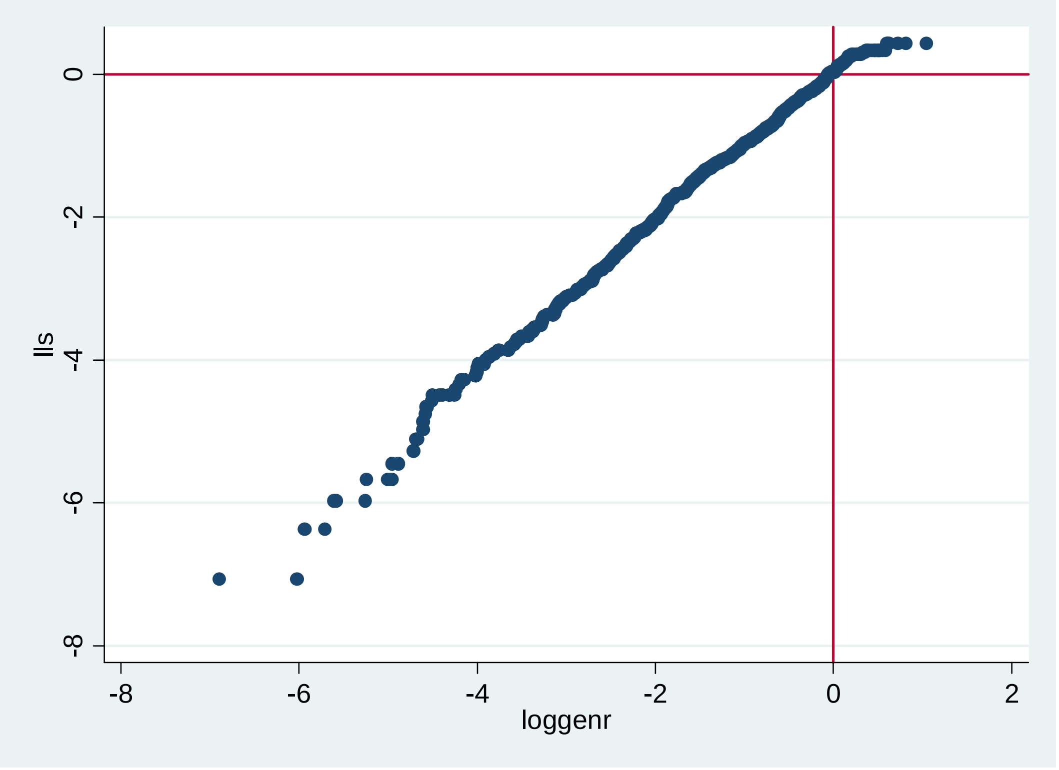
Then produce the informative graph for the generalised residuals we have to define a survival dataset using the Cox-Snell residuals as the "pseudo" failure times.

```
stset csres, failure(dthstat)

     failure event:  dthstat ~= 0 & dthstat ~= .
obs. time interval:  (0, csres]
 exit on or before:  failure

------------------------------------------------------------------------------
     1177  total obs.
        2  event time missing (csres==.)                          PROBABLE ERROR
        3  obs. end on or before enter()
------------------------------------------------------------------------------
     1172  obs. remaining, representing
      514  failures in single record/single failure data
      514  total analysis time at risk, at risk from t =         0
                             earliest observed entry t =         0
                                  last observed exit t =  2.886069
```

```
sts gen survcs=s

gen lls=log(-log(survcs))
(6 missing values generated)

gen loggenr=log(csres)
(5 missing values generated)
```

Then we want to plot :

```
scatter lls loggenr, yline(0) xline(0)
```



If the data fit the model well we would expect a straight line.

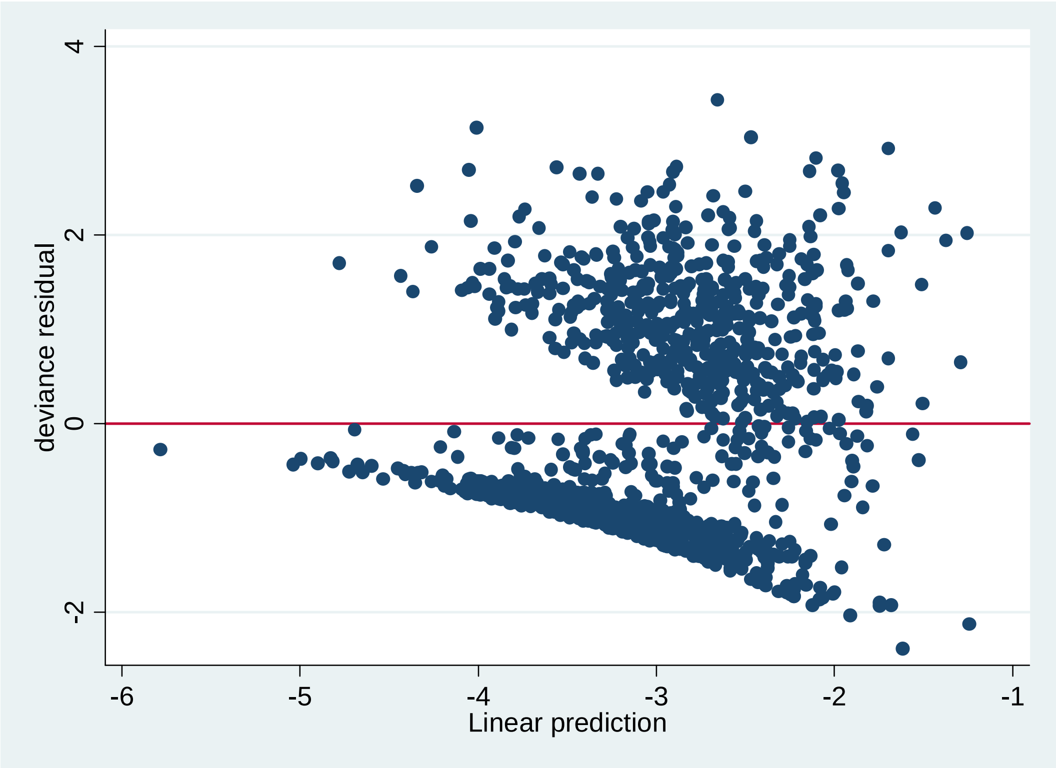
♦ **Deviance Residuals:**

Again to get the deviance residuals in STATA we first use the **stcox** command with the **mgale** option (make sure to drop **mg** before and stset the dataset using the real failure times) and then the **predict** command with the **deviance** option this time.

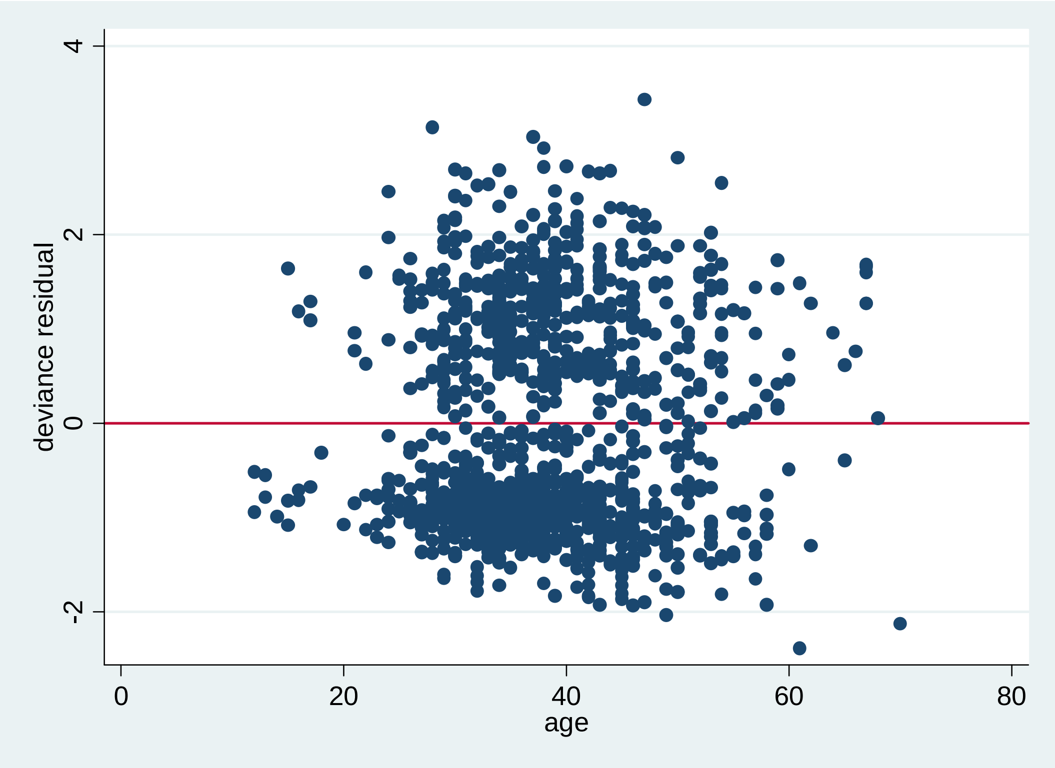
```
drop mg
stset  dthtime  dthstat

stcox  age sex cd4 karnof antiret , mgale(mg)

predict devres, deviance
(5 missing values generated)
```

And they can be plotted against the predicted log(HR) and other covariates, as shown for the Martingale residuals.

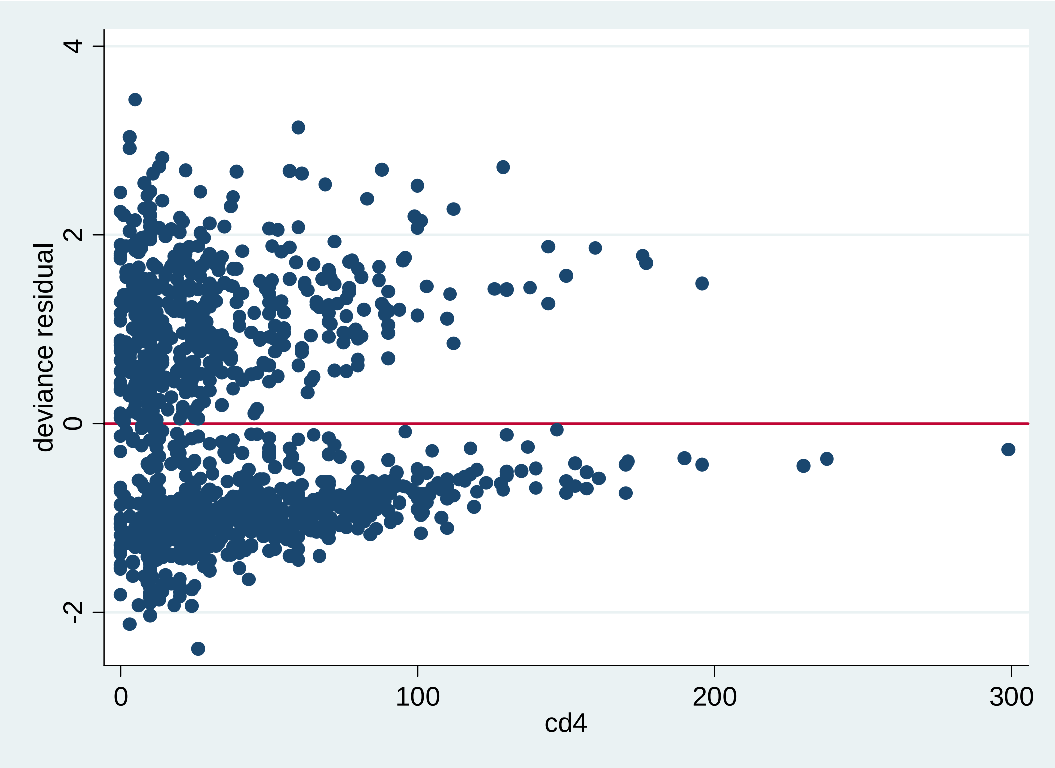
```
scatter devres betaz, yline(0)
```



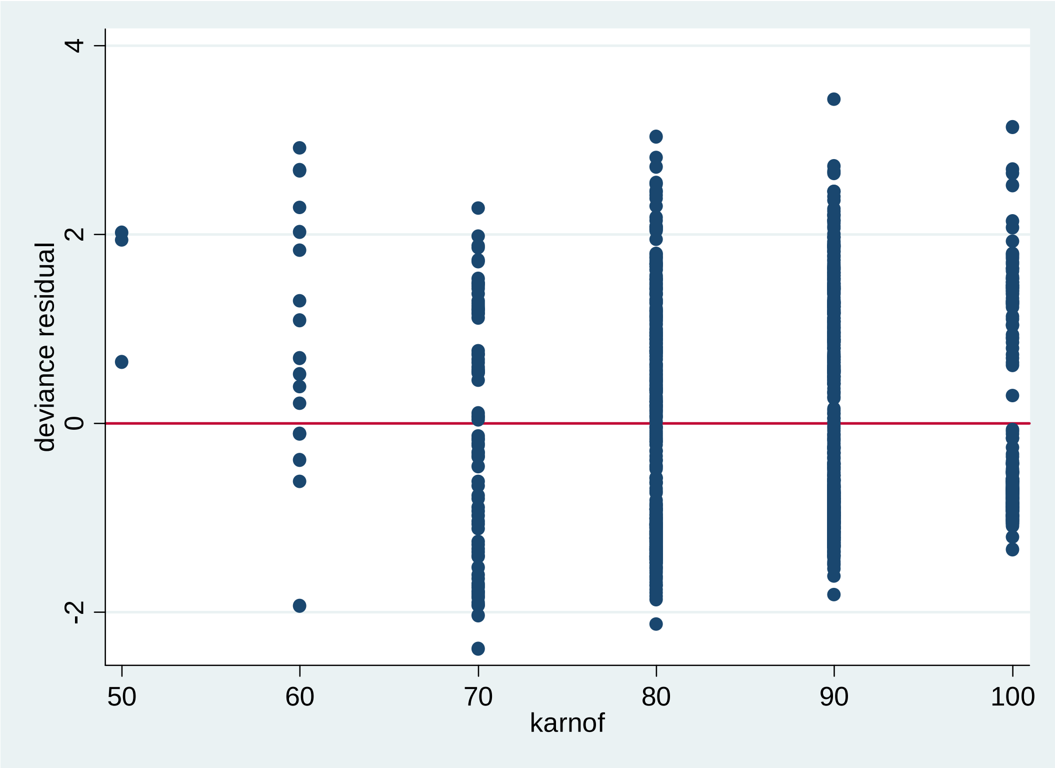
```
scatter devres age, yline(0)
```




```
scatter devres cd4, yline(0)
```



```
scatter devres karnof, yline(0)
```



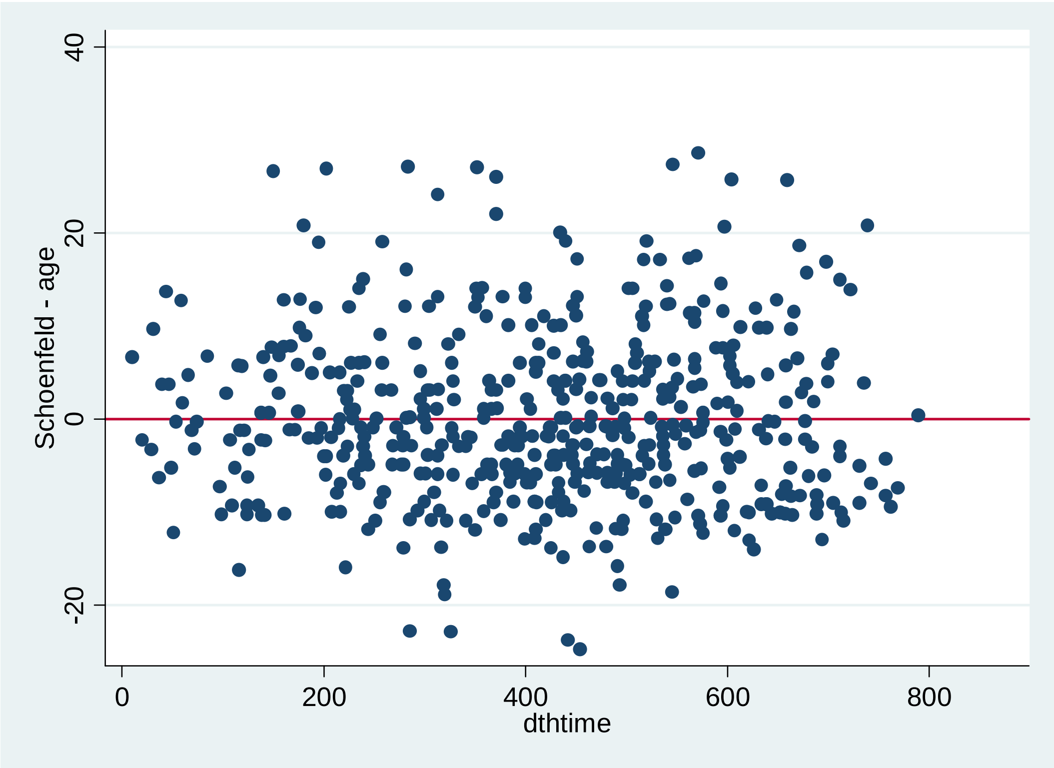
♦ **<u>Schoenfeld Residuals:</u>**

In STATA the *Schoenfeld* residuals are generated in the **`stcox`** command itself, using the **`schoenfeld (newvar(s))`** option:

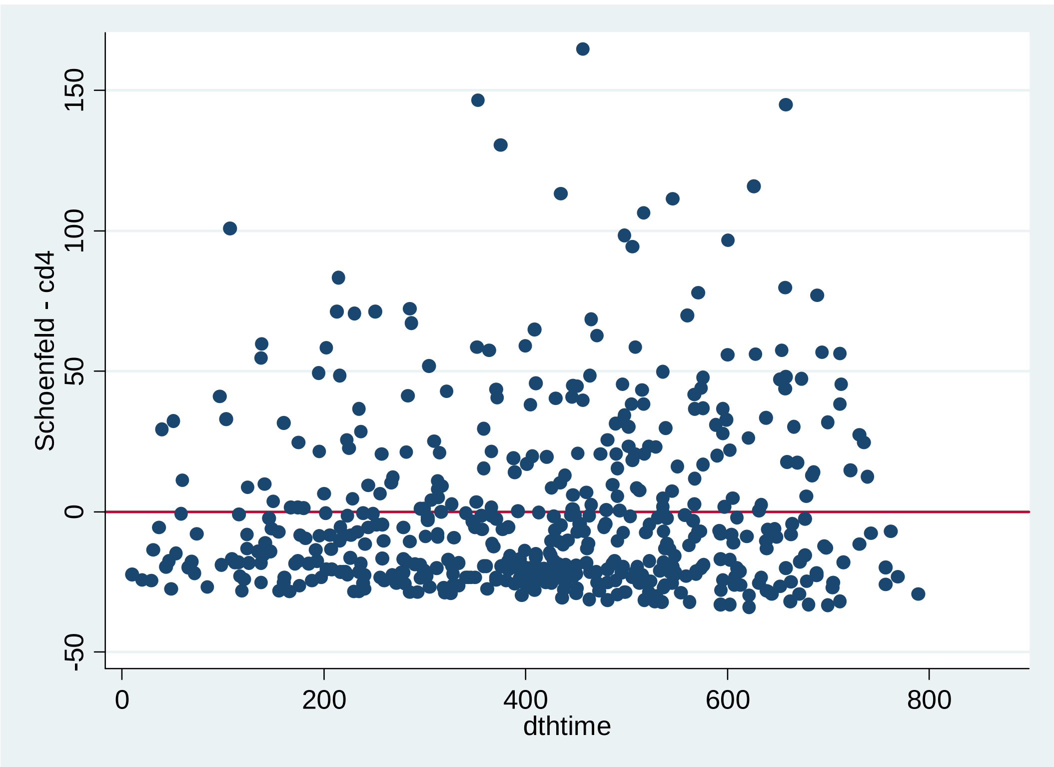
```
stcox  age sex cd4 karnof antiret , schoenf(ageres sexres cd4res
karnres antires)
 (The output is exactly the same as before)
```

Then we plot them against event time:

```
scatter ageres dthtim, yline(0)
```



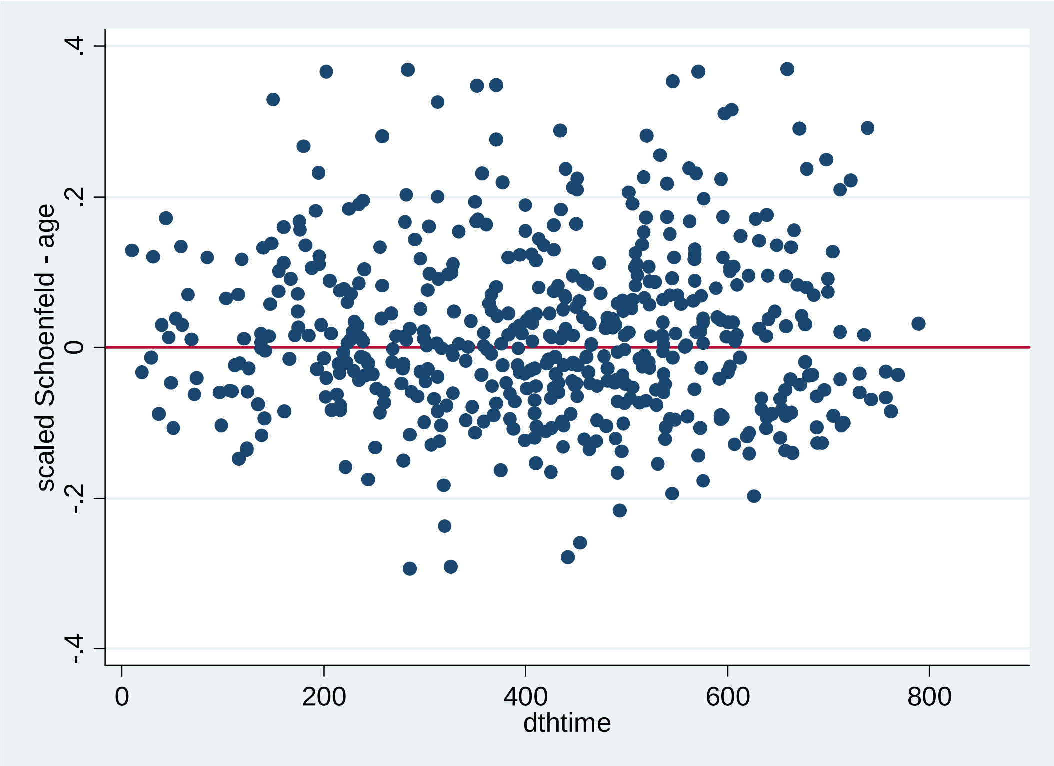

```
scatter cd4res dthtim, yline(0)
```



♦ **<u>Weighted Schoenfeld Residuals:</u>**

These residuals are used more than the previous unweighted version, because they are symmetric around 0. In this case we use the following command:

```
stcox  age sex cd4 karnof antiret , scaledsch(ageres2 sexres2 cd4res2
karnres2 antires2)
(Same output as before )

scatter ageres2 dthtim, yline(0)
```



```
scatter cd4res2 dthtim, yline(0)
```