Biostatistics in Research Practice

Exercise: Analysis of the VenUS I trial in Stata 9

In this exercise we shall explore some of the functions which Stata provides for the analysis of time to event data. The data come from the VenUS I trial of four layer elastic bandaging for venous leg ulcers. Load the file, venusi.dta, and start Stata.

Date functions

Inspect the file. You will see that there are the following variables:

1. id — Identity code
2. centre — Centre Code
3. arm — Treatment arm
4. sex — Sex
5. duration — Duration of ulcer
6. episodes — Previous episodes of ulceration
7. mobility — Mobility
8. ankcirc — Ankle circumference
9. area — Area of ulcer (sq cm)
10. age — Age
11. heal_dat — Healing date
12. entrance — Entrance date
13. last_dat — Last date

To carry out the analysis, we need two variables: the time from trial entrance to healing or censoring, and a variable which says whether the patient has healed or has been censored.

This is easy in Stata. The dates are already in date format, so we can just subtract one from another to get the time difference in days. We will first compute the difference between healing data and entrance date in days, which we will call time1. You should have:

```
gen time1 = heal_dat - entrance
```

Now repeat this to get the number of days from entrance to the last date. Make variable

```
gen time2 = last_dat - entrance
```

Have a look at **time1** and **time2**. We need to combine them to make a third time variable, which we can call **time**. This will be the time to the event (healing) or the last time seen, whichever is earlier. If there is a healing date, the patient healed and **time** = **time1**. If healing date is missing, the patient did not heal and **time** = **time2**. To do this, you can make **time** = **time1** then replace **time** by **time2** if **time1** is missing.

```
gen time = time1
replace time = time2 if time1 == .
```

Note the double "==" for "is equal to". A single "=" means "make equal to". Note the "." for missing data. I labelled this variable "**Time to healing (days)"**:

```
label var time "Time to healing (days)"
```

We could also have generated time more efficiently by

```
gen time = heal_dat – entrance
replace time = last_dat - entrance if time1 == .
```

Have a look at the dates and times. Try case number 9, Id = 1022. The dates are 10 Jun 99 to 24 Jun 99, which I would think of as a difference of 14 days. Variable **time** = 13.50269 days. I don't know why Stata does this, but we shall round them up:

```
replace time = int(time + 1)
```

to make the analysis consistent with that published. "int" is the integer part function, so for case number 9 we add 1 to give 14.50269 then chop of the fractional part to give 14.

Healed or censored?

Finally, we will need a variable for status, healed or censored. We want all non-missing **time1** subjects to have **status** = 1, healed, and all missing **time1** subjects to have **status** = 0, not healed or censored.

```
gen status = 1
replace status = 0 if time1 == .
```

Kaplan Meier survival estimates

In Stata 9, there is a special set of commands for survival analysis. They start with "st" for "survival times". To see them, try:

```
help st
```

To use them, we have to tell Stata what the survival time and outcome variables are. The command is:

```
stset time, failure(status)
```

Note the warning about "1 obs. end on or before enter()". This tells us that one observation has a zero or negative survival time. To see what this is, you could try:

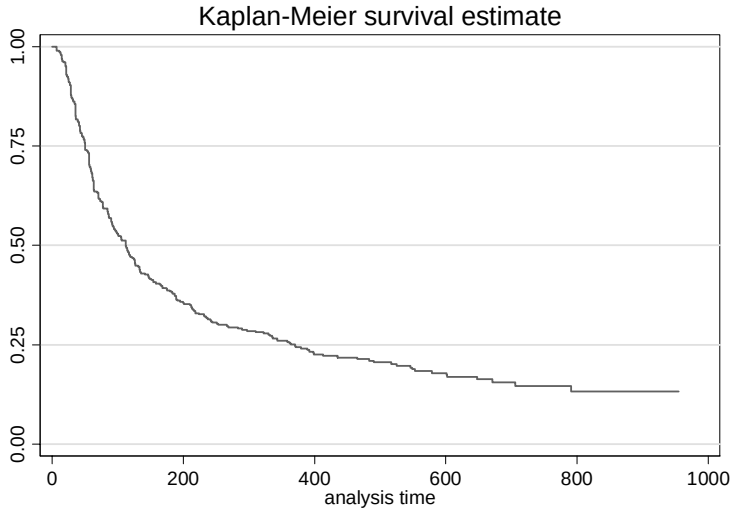
```
list if time<=0
```

This will show that case number 28, Id = 1062, was only seen on the entrance date. This subject will contribute nothing to the analysis and we can ignore it.

We can now do the Kaplan Meier analysis. This is done using the sts command. Just type st:

```
sts
```

You will get a Kaplan Meier curve for the whole dataset:



I had used

```
set scheme s1mono
```

so that I could print this. You might like

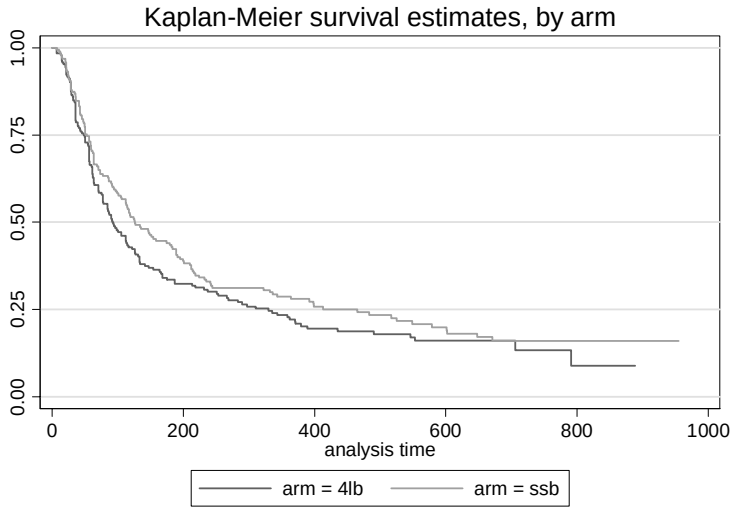
```
set scheme s1color
```

better on the screen. We get exactly the same by

```
sts graph
```

We can get the survival curves for each treatment arm by

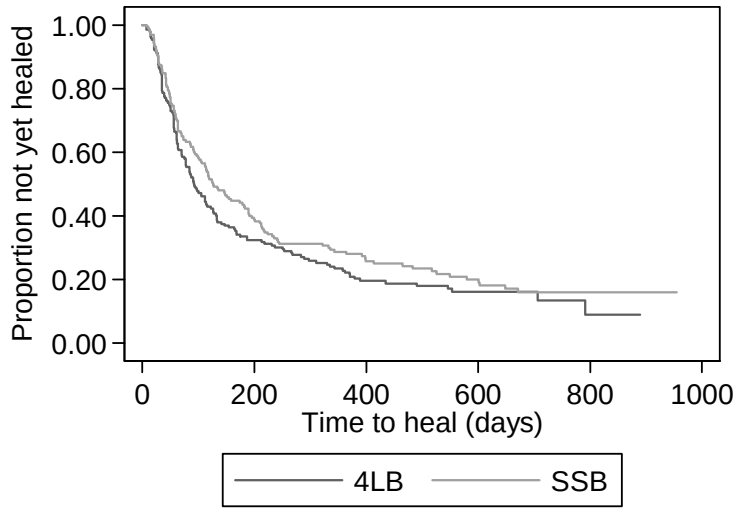
```
sts graph, by(arm)
```



This can be improved quite a lot. We can relabel the horizontal axis, label the vertical axis, change the number interval, make the numbers horizontal, remove the grid lines, remove the title, change the legend, and make the text bigger:

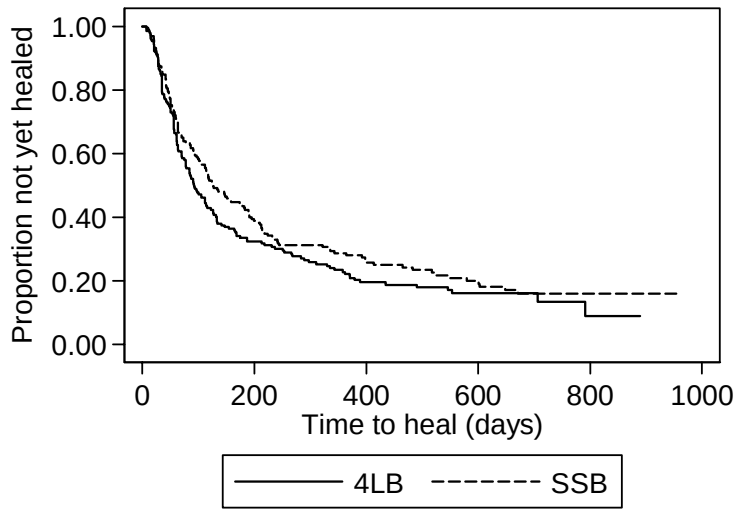
```
sts graph, by(arm) xtitle("Time to heal (days)")
ytitle(Proportion not yet healed) ylabel(0 (0.2) 1,
angle(horiz) nogrid) title("") legend(order(1 2)
label(1 "4LB") label(2 "SSB")) scale(1.5)
```

Of course, this command all goes on one line. The options are all standard graph options, apart from "by(arm)". We get:



For printing, it is often better to make one curve dashed and make them both black:

```
sts graph, by(arm) clpattern(solid dash) clcolor(black
black) xtitle("Time to heal (days)") ytitle(Proportion
not yet healed) ylabel(0 (0.2) 1, angle(horiz) nogrid)
title("") legend(order(1 2) label(1 "4LB") label(2
"SSB")) scale(1.5)
```

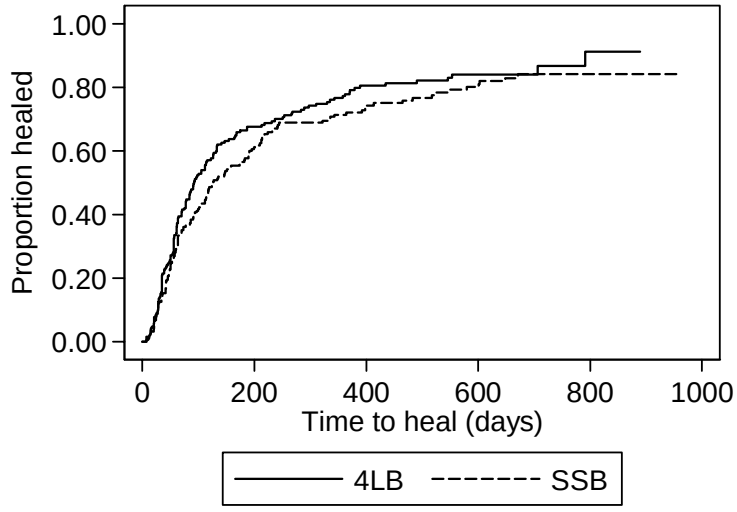


We add “clpattern(solid dash) clcolor(black black)” because the lines are connect lines (“cl”). We have to set colour and pattern for both of them. (Of course, Stata speaks American and likes “color”, not “colour”. Barbarians, or sensible, whichever you like.)

We can get the proportion healed instead of the proportion yet to heal using the “failure” option:

```
sts graph, by(arm) failure clpattern(solid dash)
clcolor(black black) xtitle("Time to heal (days)")
ytitle(Proportion healed) ylabel(0 (0.2) 1, angle(horiz)
nogrid) title("") legend(order(1 2) label(1 "4LB")
label(2 "SSB")) scale(1.5)
```

We get:



It is “failure” of course because this is called the failure function, though for us healing is a success.

Log-rank test

We can test the null hypothesis that the two treatments are the same using a log-rank test. This is also done using sts:

```
sts test arm
```

You should get:

```
. sts test arm

         failure _d:  status
   analysis time _t:  time


Log-rank test for equality of survivor functions

      |   Events         Events
arm   |  observed       expected
------+------------------------
4lb   |       154         140.86
ssb   |       144         157.14
------+------------------------
Total |       298         298.00

           chi2(1) =       2.35
           Pr>chi2 =     0.1250
```

The sts command will also give you all the survival estimates at every time, Greenwood confidence bounds, etc., as options.

Cox regression

Carry out Cox regression of survival on treatment arm and area of ulcer. You can do this by the stcox command:

```
stcox arm area
```

This gives:

```
. stcox arm area

         failure _d:  status
   analysis time _t:  time

Iteration 0:   log likelihood =  -1566.596
Iteration 1:   log likelihood = -1553.6348
Iteration 2:   log likelihood = -1548.8007
Iteration 3:   log likelihood = -1547.6302
Iteration 4:   log likelihood =  -1547.574
Iteration 5:   log likelihood = -1547.5738
Refining estimates:
Iteration 0:   log likelihood = -1547.5738

Cox regression -- Breslow method for ties

No. of subjects =          385                     Number of obs   =       385
No. of failures =          298
Time at risk    =        76522
                                                   LR chi2(2)      =     38.04
Log likelihood  =   -1547.5738                     Prob > chi2     =    0.0000
```

```
------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         arm |   .7931181   .0922428    -1.99   0.046     .6314512    .9961758
        area |   .9733582   .0062712    -4.19   0.000     .9611442    .9857274
------------------------------------------------------------------------------
```

The hazard ratio for **arm** tells us that treatment arm changing from 0 (4LB) to 1 (SSB) reduces the probability of healing on any given day by a factor 0.793. Alternatively, treatment arm changing from 1 (SSB) to 0 (4LB) increases the probability of healing on any given day by a factor 1/0.793 = 1.26. The difference is significant, just.

We could get the hazard ratio the other way up by recoding the **arm** variable to 1 for SSB and 2 for 4LB. I made a new variable, **arm2 = 2 – arm**:

```
gen arm2 = 2 - arm
stcox arm2 area
```

We get:

```
. stcox arm2 area

         failure _d:  status
   analysis time _t:  time

Iteration 0:   log likelihood =  -1566.596
Iteration 1:   log likelihood = -1553.6348
Iteration 2:   log likelihood = -1548.8007
Iteration 3:   log likelihood = -1547.6302
Iteration 4:   log likelihood =  -1547.574
Iteration 5:   log likelihood = -1547.5738
Refining estimates:
Iteration 0:   log likelihood = -1547.5738

Cox regression -- Breslow method for ties

No. of subjects =          385                     Number of obs   =       385
No. of failures =          298
Time at risk    =        76522
                                                   LR chi2(2)      =     38.04
Log likelihood  =   -1547.5738                     Prob > chi2     =    0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        arm2 |   1.260846   .1466414     1.99   0.046     1.003839    1.583654
        area |   .9733582   .0062712    -4.19   0.000     .9611442    .9857274
------------------------------------------------------------------------------
```

We can also include other variables. In a multicentre trial, it is usual to include the centre as a predictor. We can add **centre** to the covariates. Centre is categorical, so we have to create some dummy variables. In Stata, we can do this using the "xi:" (interaction expansion) command:

```
xi: stcox arm area i.centre
```

The "xi:" tells Stata to create the dummy variables and the "i." tells it for which variable these are to be created. We get:

```
. xi: stcox arm area i.centre
i.centre          _Icentre_1-9       (naturally coded; _Icentre_1 omitted)

         failure _d:  status
   analysis time _t:  time
```

```
Iteration 0:   log likelihood =  -1566.596
Iteration 1:   log likelihood = -1538.4676
Iteration 2:   log likelihood = -1531.2707
Iteration 3:   log likelihood = -1530.4901
Iteration 4:   log likelihood = -1530.4617
Iteration 5:   log likelihood = -1530.4617
Refining estimates:
Iteration 0:   log likelihood = -1530.4617

Cox regression -- Breslow method for ties

No. of subjects =          385                     Number of obs   =       385
No. of failures =          298
Time at risk    =        76522
                                                   LR chi2(10)     =     72.27
Log likelihood  =   -1530.4617                     Prob > chi2     =    0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         arm |   .7508482   .0890128    -2.42   0.016     .5951721    .9472438
        area |    .976892   .0059162    -3.86   0.000     .9653649    .9885567
  _Icentre_2 |   1.283421   .2033215     1.58   0.115     .9408501    1.750724
  _Icentre_3 |   2.465417   .4210099     5.28   0.000     1.764144    3.445457
  _Icentre_4 |   1.323886   .3058625     1.21   0.225     .8417711    2.082126
  _Icentre_5 |   1.112807   .2742678     0.43   0.665     .6864809    1.803894
  _Icentre_6 |   .9018402   .2931813    -0.32   0.751     .4768816    1.705488
  _Icentre_7 |   .7933536    .337384    -0.54   0.586     .3447325    1.825792
  _Icentre_8 |    .798684   .3410406    -0.53   0.599     .3458671    1.844339
  _Icentre_9 |   .5626886   .3321524    -0.97   0.330     .1769326    1.789486
------------------------------------------------------------------------------
```

The only estimates of interest are for **area** and **arm**. The effect is to make the hazard ratio for **arm** a bit smaller, with a lower P value, and reduce the effect of area (hazard ratio closer to 1.00). This is because area is significantly related to centre, some centres had patients with worse ulcers than others.

We can put centre in as strata rather than a categorical variable. This is an option in the stcox command:

`stcox arm area, strata(centre)`

We get a very slightly different estimate:

```
. stcox arm area, strata(centre)

        failure _d:  status
  analysis time _t:  time

Iteration 0:   log likelihood = -1040.5622
Iteration 1:   log likelihood = -1028.8549
Iteration 2:   log likelihood = -1025.4944
Iteration 3:   log likelihood = -1024.9019
Iteration 4:   log likelihood = -1024.8847
Iteration 5:   log likelihood = -1024.8847
Refining estimates:
Iteration 0:   log likelihood = -1024.8847

Stratified Cox regr. -- Breslow method for ties

No. of subjects =          385                     Number of obs   =       385
No. of failures =          298
Time at risk    =        76522
                                                   LR chi2(2)      =     31.35
Log likelihood  =   -1024.8847                     Prob > chi2     =    0.0000
```

```
------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         arm |   .7657379   .0911484    -2.24   0.025     .6063992    .9669446
        area |   .9771864   .0059442    -3.79   0.000     .9656051    .9889066
------------------------------------------------------------------------------
                                                        Stratified by centre
```

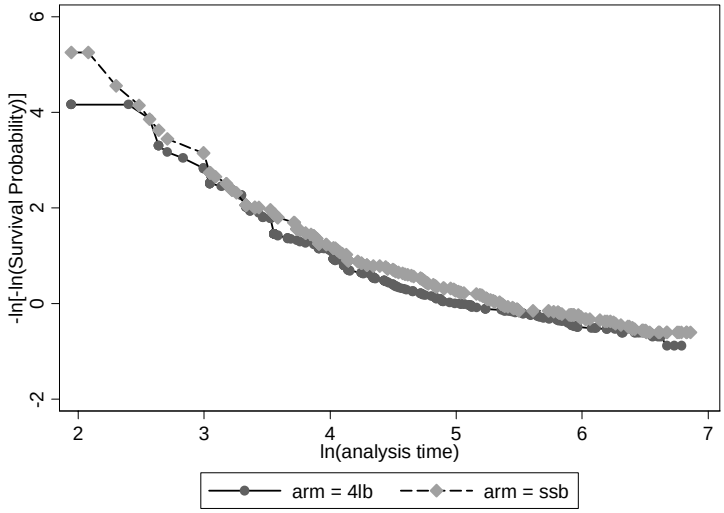
The estimate for treatment is very slightly different. I am not sure why this is, but I think that either analysis is acceptable.

Log minus log survival

And finally, let us check the assumption of proportional hazards. This is easy in Stata. We have a command "stphplot" (survival time proportional hazards plot):

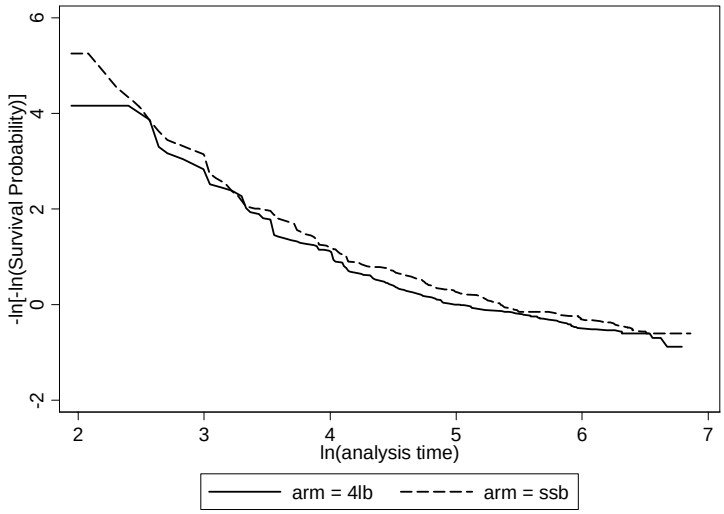
```
stphplot, by(arm)
```

We get:



We can tidy this up using all the usual graph options. For example, we can make the marker symbols invisible:

```
stphplot , by(arm) msymbol(i i)
```

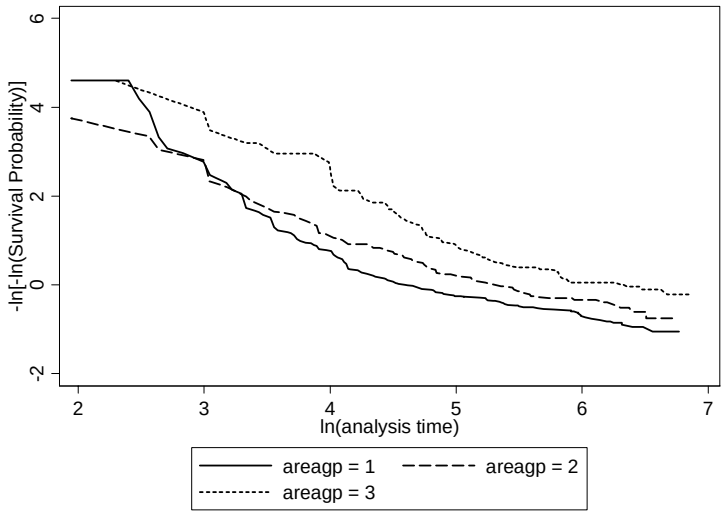


Try checking the assumption of proportional hazards for area, by creating a variable which gives are as three groups: less than 4 cm^2, 4 to 8 cm^2, more than 8 cm^2.

We get:

```
. gen areagp = 1 if area<4
(187 missing values generated)

. replace areagp = 2 if area>=4 & area<=8
(87 real changes made)

. replace areagp = 3 if area>8 & area !=.
(99 real changes made)

.
. stphplot, by(areagp) msymbol(i i i)

        failure _d:  status
  analysis time _t:  time
```

We get:



Note that second replace command include a instruction to do this only if area is not missing, because the missing data code in Stata is a very large number.

The middle line crosses the other in the same way, showing that the lines are not parallel and the proportional hazards assumption is not well met.

Checking for change in risk over time

One further point about the analysis is the checking of the assumption that early and late entrants are the same.

We need to create a variable which shows us early and late entrants. One way do this we first order the file by entrance date:

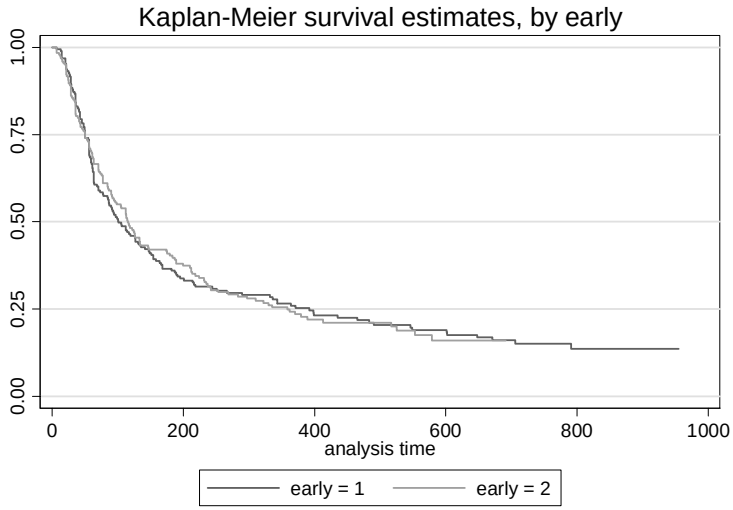
```
. sort entrance
```

We now generate a new variable:

```
. gen early = 1
. replace early= 2 if _n > 193
(194 real changes made)
```

We use 193 because there are 387 subjects, 387/2 = 193.5. Now we can draw the two curves:

```
. sts graph, by(early)

        failure _d:  status
  analysis time _t:  time
```



.

There appears to be little difference. We can check with a log rank test:

```
. sts test early

        failure _d:  status
  analysis time _t:  time


Log-rank test for equality of survivor functions

      |   Events         Events
early |  observed       expected
------+------------------------
1     |       152         151.58
2     |       146         146.42
------+------------------------
Total |       298         298.00

           chi2(1) =       0.00
           Pr>chi2 =     0.9608
```

We could also treat time as continuous (which it is) and do Cox regression:

```
. stcox entrance

        failure _d:  status
  analysis time _t:  time

Iteration 0:   log likelihood = -1567.7101
Iteration 1:   log likelihood = -1567.3585
Refining estimates:
Iteration 0:   log likelihood = -1567.3585
```

```
Cox regression -- Breslow method for ties

No. of subjects =          386                     Number of obs   =       386
No. of failures =          298
Time at risk    =        76747
                                                   LR chi2(1)      =      0.70
Log likelihood  =   -1567.3585                     Prob > chi2     =    0.4017

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    entrance |   1.000315   .0003754     0.84   0.402     .9995792    1.001051
------------------------------------------------------------------------------
```

Either way, there is no evidence for any effect.

If there is an effect, we can adjust for it by putting entrance date into the Cox model:

```
. stcox arm2 area entrance

         failure _d:  status
   analysis time _t:  time

Iteration 0:   log likelihood =  -1566.596
Iteration 1:   log likelihood = -1553.6298
Iteration 2:   log likelihood = -1548.8004
Iteration 3:   log likelihood = -1547.6302
Iteration 4:   log likelihood = -1547.5739
Iteration 5:   log likelihood = -1547.5738
Refining estimates:
Iteration 0:   log likelihood = -1547.5738

Cox regression -- Breslow method for ties

No. of subjects =          385                     Number of obs   =       385
No. of failures =          298
Time at risk    =        76522
                                                   LR chi2(3)      =     38.04
Log likelihood  =   -1547.5738                     Prob > chi2     =    0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        arm2 |   1.260748   .1474316     1.98   0.048     1.002511    1.585505
        area |   .9733601   .0062786    -4.19   0.000     .9611318     .985744
    entrance |   1.000002   .0003807     0.01   0.995     .9992565    1.000749
------------------------------------------------------------------------------
```

If there is an entrance time effect, this will improve the estimate of the hazard ratio, but the Kaplan Meier curve will still be biased.

There is an SPSS version of this exercise on my website, so you can compare and contrast.

Martin Bland
February 2008