Biostatistics in Research Practice

Exercise: Analysis of the VenUS I trial in SPSS

In this exercise we shall explore some of the functions which SPSS 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.sav, and start SPSS.

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.

We shall use the date functions for this. There are lots of these in SPSS. Go to Transform, Compute. Put in a new variable name time1. Look at the functions available. Click Date arithmetic and you will see three possibilities. We will use Datediff, so click it and put it into the Numeric Expression box. You will see the details of what it does appear. We will compute the difference between healing data and entrance date in days. You should have:

DATEDIFF(heal_dat, entrance, "days")

in the Numeric Expression box. Click OK. Now repeat this to get the number of days from entrance to the last date. Make variable

time2 = DATEDIFF(last_dat, entrance, "days").

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**. Use Transform, Compute to do this. Make **time** = **time1** and use the If condition. You can make the condition

MISSING(time1) = 0

MISSING(time1) = 1 if **time1** is missing, **MISSING(time1)** = 0 if **time1** is not missing. Then make **time** = **time2** and use the If condition:

MISSING(time1) = 1

I labelled this variable **Time to healing (days)**. Have a look at the dates and times. Try case number 7. The dates are 14-JUN-1999 to 28-JUN-1999, which I would think of as a difference of 14 days. Variable **time** = 13 days. SPSS does not include either the start or the

end date, I don't know why. In the original analysis, times were all increased by 1 day to correct this, so we shall do the same. Use Transform, Compute

time = time + 1

to make the analysis consistent with that published. (Don't forget to remove the If condition.)

Healed or censored?

Finally, we will need a variable for status, healed or censored. I did this by Transform, Recode, Into new variables, which I called **status**. I wanted all non-missing **time1** subjects to have **status** = 1, healed, and all missing **time1** subjects to have **status** = 0, not healed or censored. I recoded Old Value System-missing to have New Value 0, Add, and Old Value Range 0 through 1000000 to have New Value 1, Add, Continue, Change, OK. I put in value labels 1 = 'Healed' and 0 = 'Censored'.

Kaplan Meier survival estimates

This is easy to find in SPSS. Analysis, Survival, Kaplan Meier will do it. Put **Time to healing** into Time: and **Status** into Status:. Define Event, Single value, 1, Continue tells SPSS whether the subject has healed or is censored. We want separate plots for the treatment groups so Factor:, **Treatment arm**. We want a survival curve, so Options, Plots Survival, Continue, OK.

You should get:

Case Processing Summary

Treatment arm	Total N	N of Events	Censored	
			N	Percent
0	195	154	41	21.0%
1	192	144	48	25.0%
Overall	387	298	89	23.0%

This shows us that we had 195 4LB patients and 192 SSB patients.

Survival Table

Treatment arm		Time	Status	Cumulative Proportion Surviving at the Time		N of Cumulative Events	N of Remaining Cases
				Estimate	Std. Error		
0	1	7.000	Healed	.	.	1	193
	2	7.000	Healed	.	.	2	192
	3	7.000	Healed	.985	.009	3	191
	4	13.000	Healed	.979	.010	4	190
	5	13.000	Censored	.	.	4	189
	6	14.000	Healed	.	.	5	188
	7	14.000	Healed	.	.	6	187
	8	14.000	Healed	.964	.013	7	186

. . . etc.

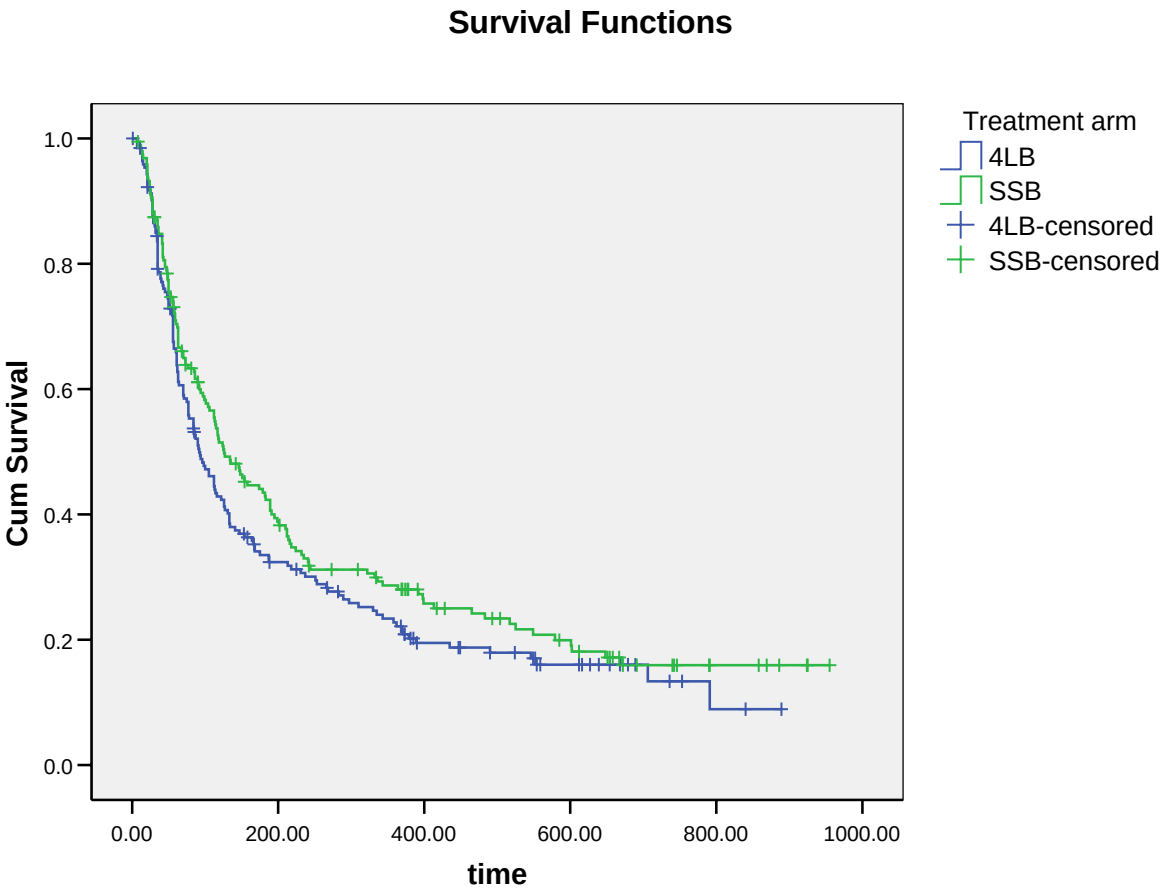
This should give you survival estimates similar to those in the lecture notes. Note that they are not exactly the same as the analysis in the notes, because of some discrepancies in the data. I have no idea which variable is correct.

Means and Medians for Survival Time

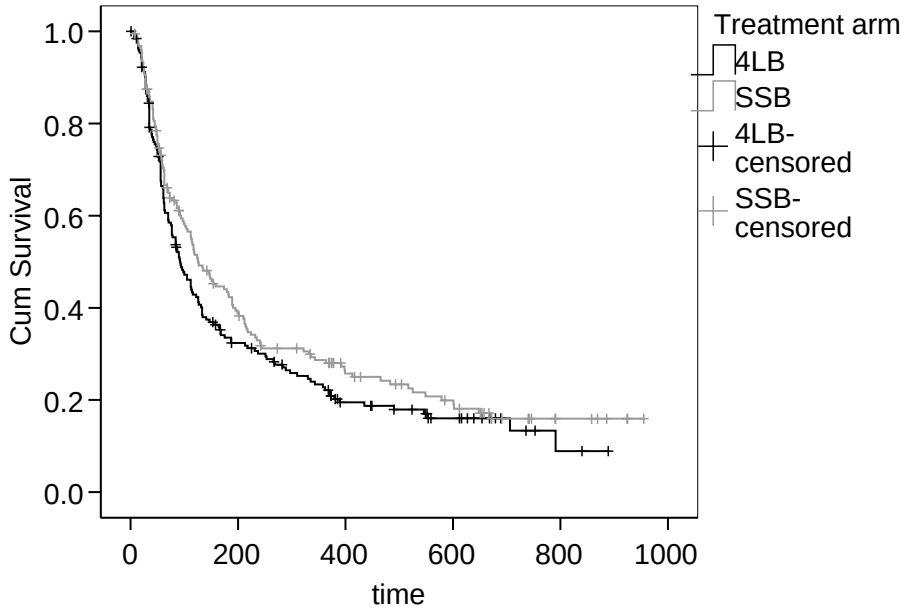
Treatment arm	Mean(a)				Median			
			95% Confidence Interval				95% Confidence Interval	
	Estimate	Std. Error	Lower Bound	Upper Bound	Estimate	Std. Error	Lower Bound	Upper Bound
0	237.916	22.030	194.737	281.094	92.000	11.452	69.554	114.446
1	292.636	25.478	242.700	342.573	126.000	15.732	95.165	156.835
Overall	269.188	17.456	234.975	303.402	112.000	8.855	94.645	129.355

a Estimation is limited to the largest survival time if it is censored.

The means are only estimates, because we do not have all the survival times, as the footnote says. The median will be correct if we have a lot of events, as we do in this data set.



Like most SPSS graphs, this benefits from a bit of editing, changing text size, number format, background to white, lines from colour to monochrome:



I have made the text larger and put the lines into black and grey. A series of slow single left clicks will pick out a particular line or a particular censored symbol so you can change the colour. (I think that if you want to produce publication quality survival curves, you should learn a different program, such as Stata.)

Log-rank test

We can test the null hypothesis that the two treatments are the same using a log-rank test. This is found in the Kaplan-Meier menu box in Compare Factor. Click Log rank, Continue. It might be a good idea to go into Options and remove all the checks, to avoid getting it all again. Then OK. You should get:

Overall Comparisons

	Chi-Square	df	Sig.
Log Rank (Mantel-Cox)	2.349	1	.125

Test of equality of survival distributions for the different levels of Treatment arm.

That is all SPSS gives you.

Cox regression

Carry out Cox regression of survival on treatment arm and area of ulcer. You can do this by Survival, Cox regression. Put variables **time** and **status** in and define the event as for Kaplan-Meier.

You should get, after some preliminaries:

Omnibus Tests of Model Coefficients(a,b)

-2 Log Likelihood	Overall (score)			Change From Previous Step			Change From Previous Block		
	Chi-square	df	Sig.	Chi-square	df	Sig.	Chi-square	df	Sig.
3096.487	21.108	2	.000	38.013	2	.000	38.013	2	.000

a Beginning Block Number 0, initial Log Likelihood function: -2 Log likelihood: 3134.500
b Beginning Block Number 1. Method = Enter

Variables in the Equation

	B	SE	Wald	df	Sig.	Exp(B)
arm	-.232	.116	3.961	1	.047	.793
area	-.027	.006	17.548	1	.000	.973

Exp(B) is the hazard ratio. It 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 can also get 95% confidence intervals for the hazard ratios in the Options box:

Variables in the Equation

							95.0% CI for Exp(B)	
	B	SE	Wald	df	Sig.	Exp(B)	Lower	Upper
arm	-.232	.116	3.961	1	.047	.793	.632	.996
area	-.027	.006	17.548	1	.000	.973	.961	.986

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**):

Variables in the Equation

							95.0% CI for Exp(B)	
	B	SE	Wald	df	Sig.	Exp(B)	Lower	Upper
area	-.027	.006	17.548	1	.000	.973	.961	.986
arm2	.232	.116	3.961	1	.047	1.260	1.004	1.583

We can also include other variables. In a multicentre trial, it is usual to include the categorical variable as a covariate. In the Cox regression box, add **centre** to the covariates. Now we have to create some dummy variables. Click Categorical and put **centre** into the Categorical Covariates box, Continue, OK.

We get a table showing how the dummy variables are coded:

Categorical Variable Codings(b)

		Frequency	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
centre(a)	1.00	108	1	0	0	0	0	0	0	0
	2.00	106	0	1	0	0	0	0	0	0
	3.00	69	0	0	1	0	0	0	0	0
	4.00	38	0	0	0	1	0	0	0	0
	5.00	24	0	0	0	0	1	0	0	0
	6.00	15	0	0	0	0	0	1	0	0
	7.00	9	0	0	0	0	0	0	1	0
	8.00	9	0	0	0	0	0	0	0	1
	9.00	8	0	0	0	0	0	0	0	0

a Indicator Parameter Coding
b Category variable: centre (Centre Code)

and this output:

Variables in the Equation

	B	SE	Wald	df	Sig.	Exp(B)	95.0% CI for Exp(B)	
							Lower	Upper
area	-.023	.006	14.895	1	.000	.977	.965	.989
arm2	.286	.119	5.811	1	.016	1.331	1.055	1.679
centre			37.249	8	.000			
centre(1)	.578	.590	.959	1	.327	1.782	.561	5.668
centre(2)	.827	.589	1.973	1	.160	2.288	.721	7.258
centre(3)	1.480	.594	6.215	1	.013	4.394	1.372	14.067
centre(4)	.857	.614	1.949	1	.163	2.355	.708	7.839
centre(5)	.686	.619	1.228	1	.268	1.985	.590	6.675
centre(6)	.476	.652	.532	1	.466	1.609	.448	5.775
centre(7)	.344	.709	.236	1	.627	1.410	.352	5.655
centre(8)	.351	.708	.246	1	.620	1.420	.355	5.687

The only estimates of interest are for **area** and **arm2**. The effect is to make the hazard ratio for **arm** a bit bigger, 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. We get a very slightly different estimate:

Variables in the Equation

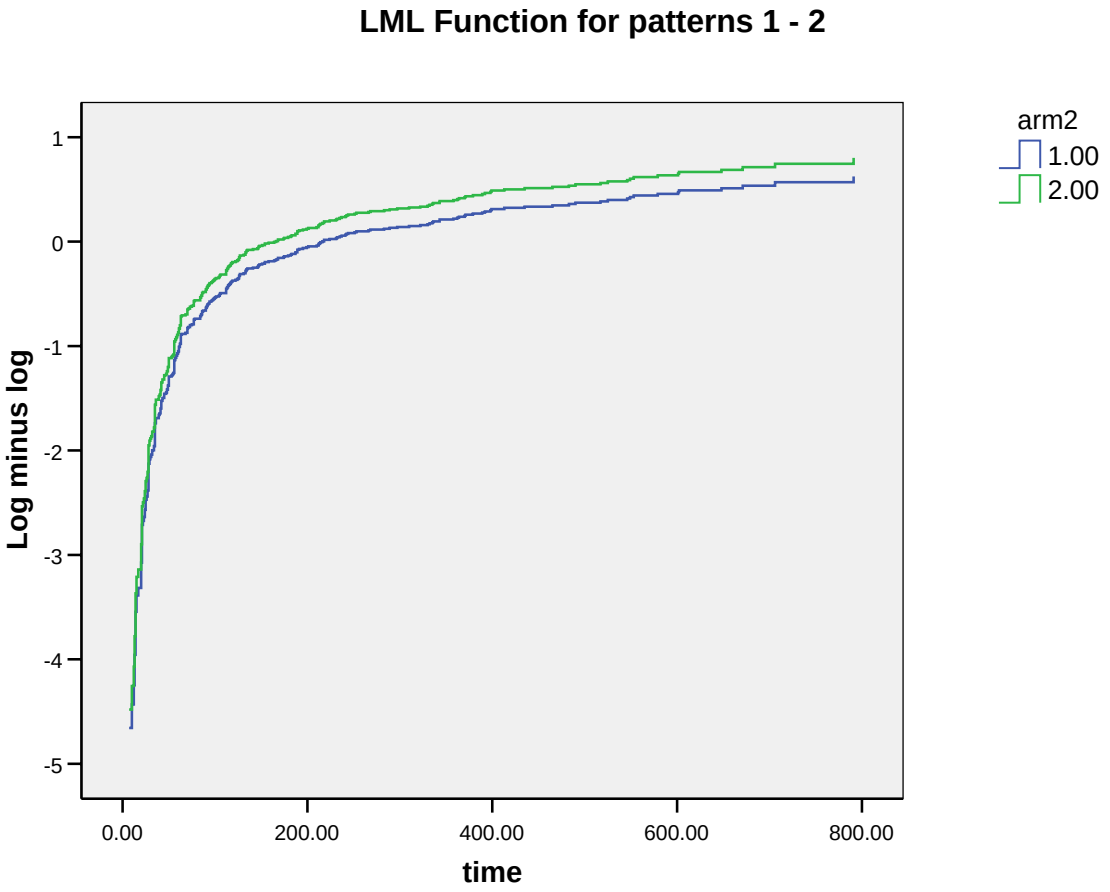
	B	SE	Wald	df	Sig.	Exp(B)	95.0% CI for Exp(B)	
							Lower	Upper
area	-.023	.006	14.421	1	.000	.977	.966	.989
arm2	.269	.119	5.109	1	.024	1.309	1.036	1.652

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

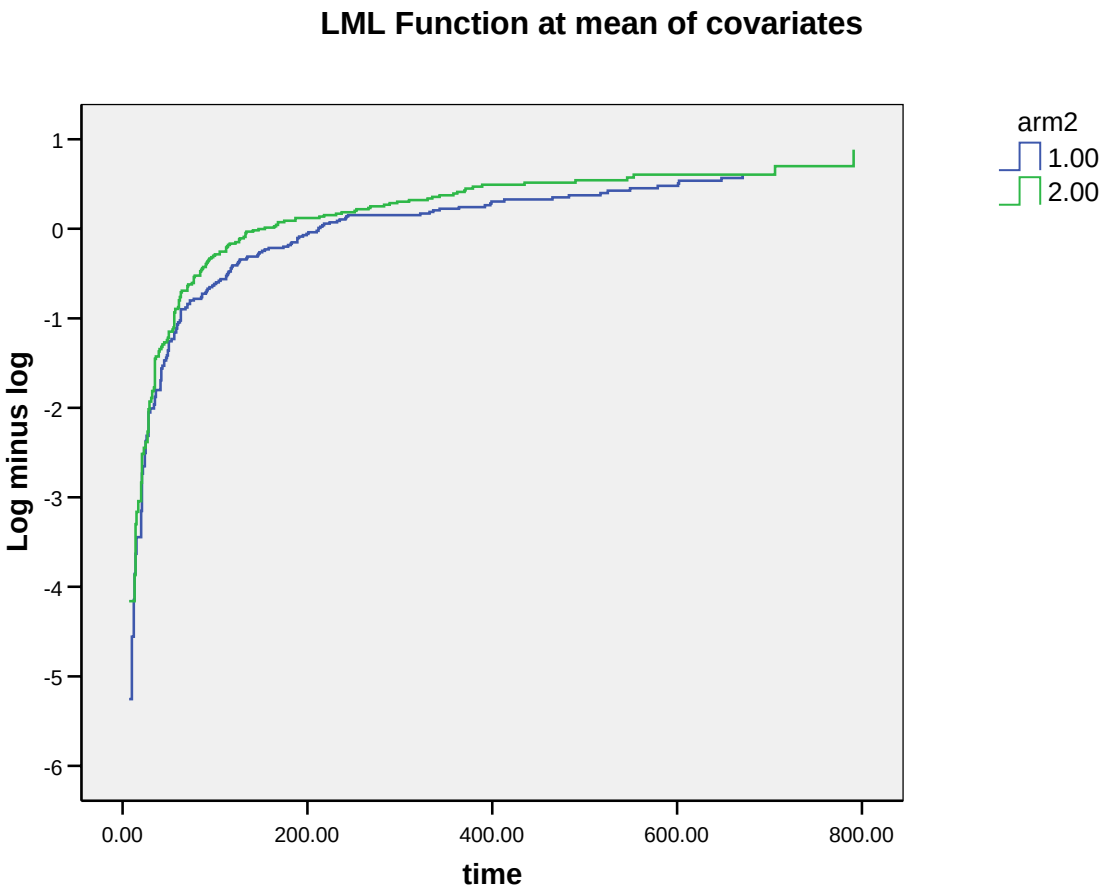
And finally, let us check the assumption of proportional hazards. As we shall see, this is a bit tricky in SPSS. For treatment group, we will use the Cox regression on **arm2**. We won't stratify by centre, as this produces a different plot for every stratum, not very useful.

The obvious thing to do appears to be to make **arm2** a categorical variable (which it is, of course). Click Categorical and put **arm2** into the Categorical Covariates box, Continue. Now we go to Plots, click Log minus log, and put **arm2(cat)** into Separate Lines for. Click Continue, OK. We get:

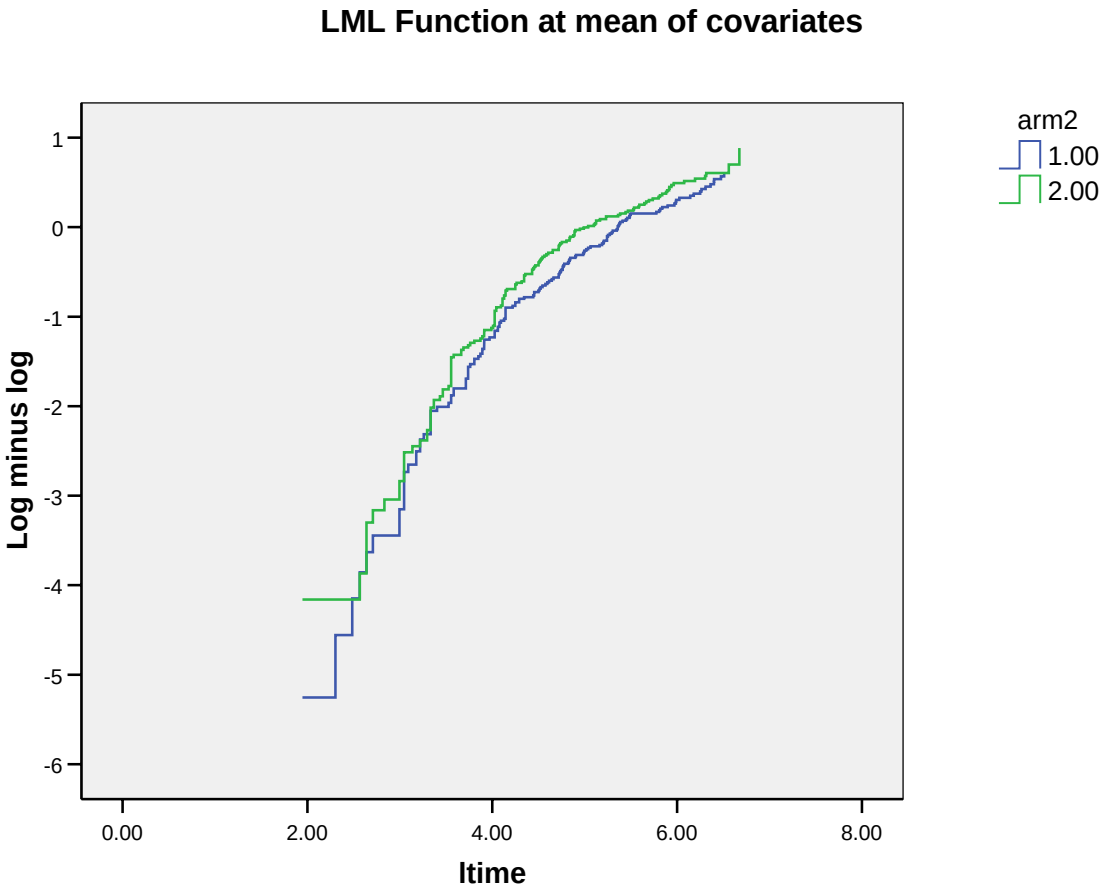


If you examine this, you will see something very odd about it. Each line has steps in exactly the same place. These steps happens whenever there is an event in ***either*** group. What SPSS is doing is producing the log minus log plot as it would be predicted by the regression model, i.e. as if there really were proportional hazards. This is ***not*** what we want. We cannot use something produced assuming proportional hazards to test whether we have proportional hazards.

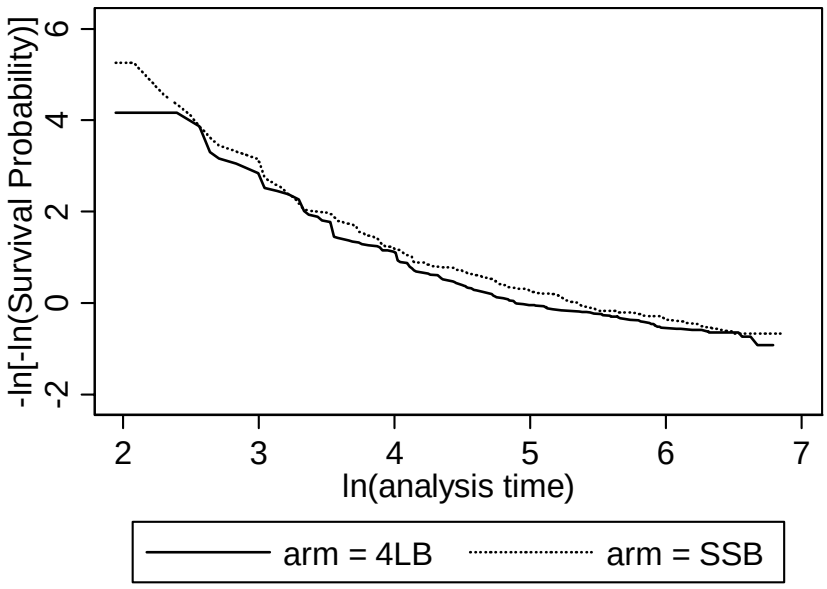
Go back to the Cox regression command and remove **arm2** from the Covariates: box and put it instead into the Strata: box, click OK. We get this:



These are clearly now different curves, with steps in different places. It is difficult to see whether they are parallel without a log scale for time. You can do this by creating a new variable, **ltime**, which is LN(**time**), i.e. the logarithm of time. Now do the Cox regression with **ltime** as the time variable instead of **time**. You should find that the various numbers are exactly the same, but the log minus log graph is a bit different:



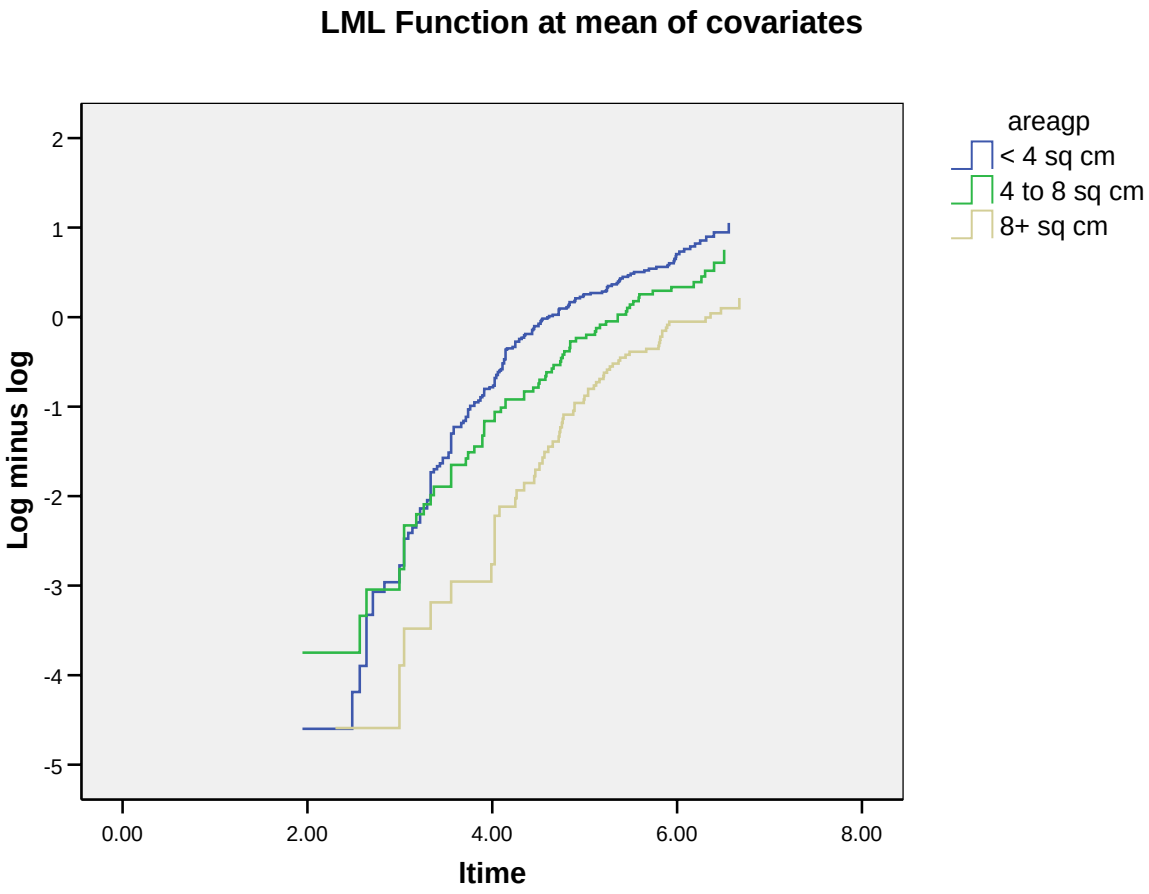
Compare this with the Stata plot:



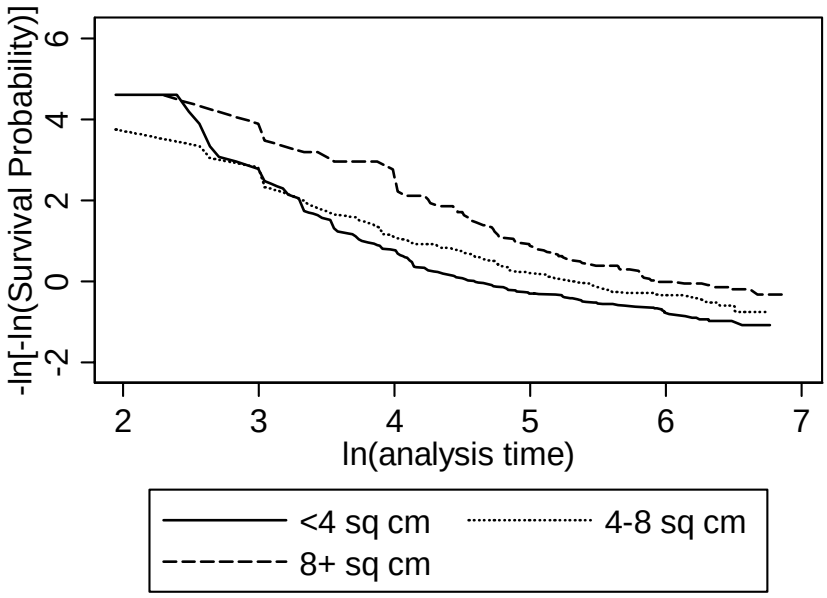
Although the Stata plot is the other way up, being for minus log minus log, rather than log minus log, and is not in steps, the pattern is the same.

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:



Compare the SPSS plot with the Stata plot:



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.

Martin Bland
20 April 2007

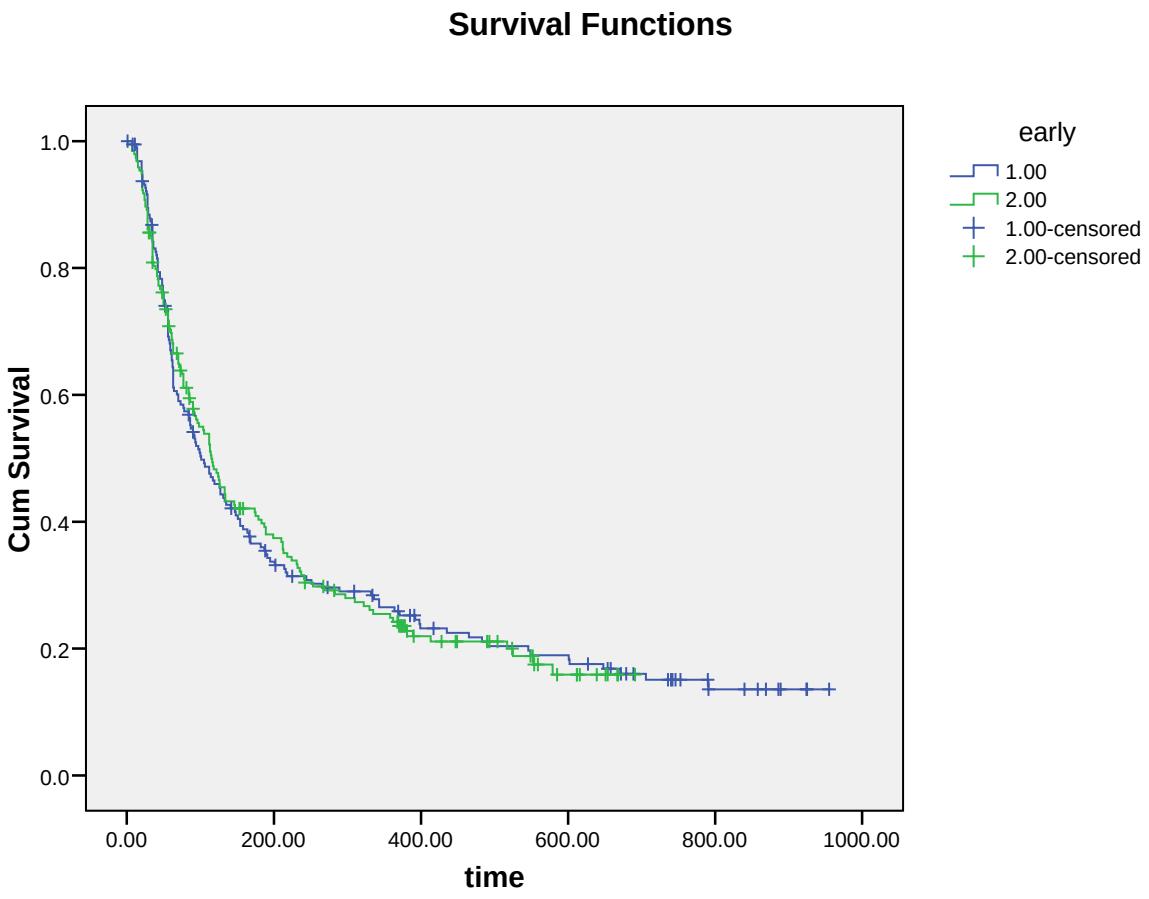
Checking for change in risk over time

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 is to first order the file by entrance date: Data, Sort cases, entrance.

We now generate a new variable: Transform, Compute Variable, new name early = 1. Then make all the later case have early = 2 by Transform, Compute Variable, early = 2, If, Include if case satissfies condition, $CASENUM > 193. We use 193 because there are 387 subjects, 387/2 = 193.5. $CASENUM is the casenumber in the current order they are in the memory. If you click All in Function Group it is the first keyword to appear. Then Continue, OK, and Change existing variable OK gets you there. Have a look to check that your new variable does what you expect.

Now we can draw the two curves survival curves using early as the factor:



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

Case Processing Summary

early	Total N	N of Events	Censored	
			N	Percent
1.00	193	152	41	21.2%
2.00	194	146	48	24.7%
Overall	387	298	89	23.0%

Overall Comparisons

	Chi-Square	df	Sig.
Log Rank (Mantel-Cox)	.002	1	.961

Test of equality of survival distributions for the different levels of early.

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

Variables in the Equation

	B	SE	Wald	df	Sig.	Exp(B)
entrance	.000	.000	.703	1	.402	1.000

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:

Variables in the Equation

	B	SE	Wald	df	Sig.	Exp(B)	95.0% CI for Exp(B)	
							Lower	Upper
entrance	.000	.000	.000	1	.995	1.000	1.000	1.000
area	-.027	.006	17.522	1	.000	.973	.961	.986
arm2	.232	.117	3.926	1	.048	1.261	1.003	1.586

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.

Martin Bland
February 2008