MAS3311/MAS8311 Biostatistics
Survival Data Analysis
Project

M. Farrow
Newcastle University

Semester 2, 2009-10

1 Background

You are to analyse some data which were collected in two studies concerning morbidity and risk factors after coronary artery bypass graft (CABG) surgery at the Freeman Hospital in Newcastle. The first study involved patients who were given grafts between January 1980 and June 1987. The second study involved patients whose operations took place between June 1987 and December 1992. For the purpose of this project the two data sets have been merged. All patients were male because surgery of this type is much more common among men and the number of female patients available for the study was rather small.

The aim of the studies was to look for associations between various potential risk factors and outcomes such as the return of angina or breathlessness and the degree of these problems. In this project you will be looking at only one response variable, the time from the operation until the onset of angina. As well as the obvious potential risk factors such as smoking, obesity and (lack of) exercise, three different surgical techniques were used and it is of interest to see whether there is evidence that these are associated with different patterns of post-operative outcomes. The three techniques were:

- Venous graft.
- Single mammary artery graft.
- Bilateral mammary artery graft.

The data were collected by means of a questionnaire given to the patient's general practitioner and by a visit to the patient. Of course, at the time that this was done, possibly several years after the operation, some patients had left the area or were untraceable and some had died. Such cases were excluded from the study.

If you would like further information on the background to the study you may ask me. You might also consult other sources of information such as the Library if you so wish.

2 Data

For Part 1 of this project you will use a small subset of the data, given in 3.1 below. The data for Parts 2 and 3 are held in two plain files:

```
cabbage1.dat
cabbage2.dat
```

The first data file contains some of the variables and the second data file contains the rest. (Actually, in the original study, there was a third file but we will not be using the variables contained in it).

You will only be using some of the variables. A R function is provided to read the data and provide you with just the required variables. In order for each student to have different data, the R function will give you a subset of the cases. Each student will get a different subset, though the subsets will overlap. Each student will have 350 cases. In order to determine which subset you get you must supply a reference number as an argument to the function. A list of reference numbers is given in Section 5 below.

When you submit Parts 2 and 3 you must confirm which reference number you have used, clearly, at the beginning of your report.

The function to read the data is called `cabread`. You can download it from the course Web page. For example, you can type the command

```
source("http://www.mas.ncl.ac.uk/~nmf16/teaching/mas3311/cabread.r")
```

in R (while using a University computer or while connected to the Internet). Once you have the function (and while still using a University computer or connected to the Internet), if your refernce number is 99, for example, you can type

```
cabbage<-cabread(99)
```

and you should obtain your data in a data frame called `cabbage`.

There were originally seven response variables:

- Degree of angina.
- Degree of breathlessness.
- Number of heart attacks since operation.
- Heart failure.
- Time in months after operation until onset of angina.
- Time in months after operation until onset of breathlessness.
- Time in months until postoperative heart attack.

This project is concerned only with the time in months after the operation until the onset of angina. Ignore the other six.

In your data set you will find the following variables.

- `cabbage$sn` the patient serial number.
- `cabbage$t` the time in months from the operation until the onset of angina or until censoring. Some observations are right-censored. The value of `cabbage$t` for a censored observation is calculated by the program using the dates of the operation and the questionnaire.
- `cabbage$status`. This indicator is 1 for a case where the onset of angina occurs and 0 for a censored observation.
- `cabbage$optype`. This is a factor with three levels, for the three operation types, as follows.

  1 : Venous graft.

  2 : Single mammary artery graft.

  3 : Bilateral mammary artery graft.
- `cabbage$ageatop` the age in years at the time of the operation.
- `cabbage$ca`. This is a factor with three levels indicating change of activity since the operation, as follows.

0 : No change,

1 : Increase,

2 : Decrease.

- `cabbage$tpreosm` total kg of tobacco smoked before the operation.
- `cabbage$preoysm` number of years smoked in five years before the operation.
- `cabbage$tpostosm` total kg of tobacco smoked since the operation.
- `cabbage$postoysm` number of years smoked in all time since the operation.
- `cabbage$enow`. This is a factor indicating exercise grade at the time of the questionnaire. The levels are the exercise grades (0,1,2,3,4).

3 Tasks

You are to submit your work in three parts as follows.

3.1 Part 1

The following table gives a small subset of the data.

Operation Type 1: Venous Graft (10 patients)		Operation Type 2: Single Mammary Artery Graft (10 patients)		Operation Type 3: Bilateral Mammary Artery Graft (20 patients)	
Time	Status	Time	Status	Time	Status
1	1	52	1	9	0
111	0	69	0	47	0
12	1	11	1	44	0
65	1	1	1	28	1
84	1	59	0	34	0
60	1	12	1	6	1
92	0	51	1	39	0
1	1	79	0	37	0
134	0	71	0	50	0
1	1	24	1	38	0
				43	0
				3	1
				24	1
				38	0
				51	0
				9	1
				47	0
				15	1
				35	0
				48	0
Status: 1 for event, 0 for censored.					

In Part 1 of the project you are to work by hand and pocket calculator and not use R. You should show enough working to make clear what you have done. Using only the data in the small subset in the table:

1. Find and plot the Kaplan-Meier estimate of the survivor function using all of the data, ignoring operation type.

2. Find and plot the Kaplan-Meier estimate of the hazard function using all of the data, ignoring operation type.

3. Find and plot the Kaplan-Meier estimate of the survivor function separately for each operation type. Plot the three estimated survivor functions on the same graph.

4. Use a log-rank test (simplified form) to test for differences between the survival with the three operation types.

You should submit your solution to this part before 4.00pm on **Tuesday 16th March**.

3.2 Part 2

In Part 2 you should use R to do calculations and plot graphs and use your large data set of 350 cases.

1. Calculate and plot the Kaplan-Meier estimate of the survivor function using all of the data, ignoring operation type.

2. Find and plot the Kaplan-Meier estimate of the survivor function stratified by operation type. Plot the three estimated survivor functions on the same graph.

3. Compare the three survivor functions using the `survdiff` function.

4. Explain what you have done and comment on the results.

You should submit your solution to this part before 4.00pm on **Tuesday 27th April**.

3.3 Part 3

In Part 3 you should use R to do calculations and plot graphs and use your large data set of 350 cases.

1. Fit a Weibull proportional hazards model to the data including the factors `optype`, `ca` and `enow` and the variables `ageatop`, `tpreosm`, `preoysm`, `tpostosm` and `postoysm`. Tabulate the parameter estimates and their standard errors.

2. Test each of the variables and factors one at a time and report the results of the significance tests.

3. On the basis of testing the variables and factors, suggest a model which contains only those variables and factors where inclusion is supported by significant evidence, fit this model and report your results. Show how you have come to this conclusion.

4. You may wish to try including extra variables such as interaction effects.

5. Present the results of your analysis (all three parts) in a report. You should include in the report an explanation of your conclusions suitable for medical readers who are not necessarily familiar with all aspects of the statistical methods you have used. Your report should also introduce the background of the data. It should be illustrated by suitable graphs and tables.

   Your report should be clear and well presented. It should not be longer than six or seven sides of A4 paper (not counting the front cover).

NOTE : Some of the times till onset of angina are recorded as 0. This is because some patients experienced angina straight away after the operation. That is, the operation was not successful in eliminating the angina even for a short time. This means that you can not fit a Weibull distribution to the data as they are. We could fit a more sophisticated model, beyond the scope of this course. However there are two simpler possibilities.

1. Remove the cases with zero times. You can do this by typing the following R command.

```
cabbage<-cabbage[cabbage$t>0,]
```

It also would be possible to build a separate model to see how the probability of a zero time depends on the covariates and factors, including operation type. You can try this if you wish but it is an optional extra.

2. Change all of the zero times to a small positive value, such as 0.1. You can do this by typing the following R command.

```
cabbage$t<-ifelse((cabbage$t==0),0.1,cabbage$t)
```

Either of these approaches will be accepted for this project. In either case you will need to form your survival object again. You might like to comment on your choice of method.

You should submit your report before 4.00pm on **Tuesday 4th May**. You will need to hand in your report as a "project" and sign at the General Office.

4 Marking

The marking scheme will be as follows.

Part 1	20%.
Part 2	15%.
Part 3 (Calculations and results)	25%
Additional ideas, comments etc. showing initiative or insight.	10%.
Clear statement of conclusions.	10%
Presentation	20%

A successful and largely correct analysis giving some usable results and presented well enough to be understood reasonably clearly will merit at least a mark of at least 50%. To obtain a mark of over 70% overall the analysis needs to be thorough in all stages, well and clearly presented with a statement of conclusions suitable for a non-expert reader and with appropriate discussion and criticism.

5 Data Reference Numbers

Badi	Nuri H.	1	Margrie-Rouse	Joseph S.	28
Barker	Lucy	2	Massaquoi	Daana	29
Bell	Matthew J.	3	Maynard	Lauren	30
Boreman	Alice K.	4	McGarel	Nicola M.	31
Briggs	Gillian	5	Pg Osman	Dk N. Nadzirah	32
Burke	Aisling B.	6	Price	Lauren	33
Chadderton	Rebecca L.	7	Riley	Lucy	34
Crocker	Pablo D.	8	Roberts	Katherine.	35
Duke	John M.	9	Smith	Graham C.	36
Francis	Bethany J.	10	Smith	Rebecca I.	37
Gosnell	Matthew	11	Speight	Lauren	38
Gregory	Samantha C.	12	Spencer	Luke	39
Hartley	Lorraine A.	13	Starr	Craig A.	40
Hartshorn	George	14	Stephenson	Lane J.	41
Herd	Christopher J.	15	Storey	Claire L.	42
House	Catherine	16	Stothard	Rachel	43
Islam	Jabirul	17	Thorpe	Luke	44
Jhoree	Lakshna T.	18	Tibbles	Gwendolyn M.	45
Johnson	Sarah V.	19	Wyllie	Samuel R.	46
Joy	Laura	20	Young	Alexander T.I.S.	47
Kozlowski	Michael J.	21	Jamison	Deborah	48
Lai	Vincent Tsz Hin	22	Mann	Kay D.	49
Laker	Sarah L.	23	Millman	Jill F.	50
Maguire	Alexandra M.	24	Nichols	Ben	51
Maisey	Will	25	Riley	Anthony D.	52
		26	Walker	Lydia K.	53
		27	Yin	Peng	54