Survival analysis

- So far, have seen:
  - response variable counted or measured (regression)
  - response variable categorized (logistic regression)

  and have predicted response from explanatory variables.
- But what if response is time until event (eg. time of survival after surgery)?
- Additional complication: event might not have happened at end of study (eg. patient still alive). But knowing that patient has “not died yet” presumably informative. Such data called *censored*.
- Enter *survival analysis*, in particular the “Cox proportional hazards model”.
- Explanatory variables in this context often called *covariates*.

Example: still dancing?

- 12 women who have just started taking dancing lessons are followed for up to a year, to see whether they are still taking dancing lessons (or have quit).
- This might depend on:
  - a treatment (visit to a dance competition)
  - woman's age (at start of study).
- Data:

Months	Dancing	Treatment	Age
1	1	0	16
2	1	0	24
2	1	0	18
3	0	0	27
4	1	0	25
5	1	0	21
11	1	0	55
7	1	1	26
8	1	1	36
10	1	1	38
10	0	1	45
12	1	1	47

About the data

- `months` and `dancing` are kind of combined response:
  - `Months` is number of months a woman was actually observed dancing
  - `dancing` is 1 if woman quit, 0 if still dancing at end of study.
- Treatment is 1 if woman went to dance competition, 0 otherwise.
- Want to do predictions for probabilities of still dancing after 3, 6, 9, 12 months for treatment group and control group, for women of ages 25 and 45.

Doing predictions

Add to data file:

```
3 . 0 25
6 . 0 25
9 . 0 25
12 . 0 25
...
3 . 1 45
6 . 1 45
9 . 1 45
12 . 1 45
```

Gives predicted survival probabilities for 3, 6, 9 and 12 months for (a) woman aged 25 in control group, (b) women aged 45 in traatment group (do other age/treatment combos also).

Censoring variable missing for these: won't affect analysis.

The code

```
data dancers;
  infile "survival1.dat";
  input months dancing treatment age;

proc phreg;
  model months*dancing(0) = age treatment;
  output out=fred survival=s;

proc print data=fred;
```

- Nothing new in reading data.
- Note specification of model: includes both survival time and censoring variable in response, and indication of what value means “censored”.
- As ever, predictions saved in output data set, then printed.

The output, edited

Model Information

Data Set	WORK.DANCERS
Dependent Variable	months
Censoring Variable	dancing
Censoring Value(s)	0
Ties Handling	BRESLOW

Number of Observations Read	28
Number of Observations Used	12

Summary of the Number of Event and Censored Values

Total	Event	Censored	Percent Censored
12	10	2	16.67

Output part 2

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	21.0016	2	<.0001
Score	14.2093	2	0.0008
Wald	5.5556	2	0.0622

Analysis of Maximum Likelihood Estimates

Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
age	1	-0.35284	0.14973	5.5532	0.0184	0.703
treatment	1	-4.28283	2.54084	2.8412	0.0919	0.014

- Overall model seems significant.
- Survival depends on age but not apparently on treatment (could be small size of data set or confounding of treatment with age).

Predicted survival probs

Obs	months	dancing	treatment	age	s
13	3	.	0	25	0.87856
14	6	.	0	25	0.56647
15	9	.	0	25	0.00000
16	12	.	0	25	0.00000
17	3	.	1	25	0.99821
18	6	.	1	25	0.99219
19	9	.	1	25	0.00000
20	12	.	1	25	0.00000
21	3	.	0	45	0.99989
22	6	.	0	45	0.99951
23	9	.	0	45	0.14589
24	12	.	0	45	0.00000
25	3	.	1	45	1.00000
26	6	.	1	45	0.99999
27	9	.	1	45	0.97378
28	12	.	1	45	0.08223

Conclusions from predicted probs

- Older women more likely to be still dancing than younger women (compare “profiles” for same treatment group).
- Effect of treatment seems to be to increase prob of still dancing (compare “profiles” for same age for treatment group vs. not)
- Would be nice to see this on a graph.

Another way of doing predictions

Instead of adding lines to data file and creating an output data set, use `baseline` command like this:

```
data dancers;
  infile "survival1.dat";
  input months dancing treatment age;

data mypred;
  input treatment age;
  datalines;
  0 25
  0 45
  1 25
  1 45
;

proc phreg data=dancers;
  model months*dancing(0) = age treatment;
  baseline out=fred covariates=mypred survival=s lower=lcl upper=ucl /
nomean;

proc print data=fred;
```

Results, including CIs

Obs	age	treatment	months	s	lcl	ucl
1	25	0	0	1.00000	.	.
2	25	0	1	0.96633	0.90266	1.00000
3	25	0	2	0.79225	0.60826	1.00000
4	25	0	4	0.63726	0.35919	1.00000
5	25	0	5	0.14748	0.05834	0.37282
6	25	0	7	0.00000	0.00000	1.00000
7	25	0	8	0.00000	0.00000	1.00000
8	25	0	10	0.00000	0.00000	1.00000
9	25	0	11	0.00000	0.00000	1.00000
10	25	0	12	0.00000	.	.
11	45	0	0	1.00000	.	.
12	45	0	1	0.99997	0.99980	1.00000
13	45	0	2	0.99980	0.99895	1.00000
14	45	0	4	0.99961	0.99760	1.00000
15	45	0	5	0.99835	0.99486	1.00000
16	45	0	7	0.75954	0.52629	1.00000
17	45	0	8	0.04468	0.00002	1.00000
18	45	0	10	0.00001	0.00000	1.00000
19	45	0	11	0.00000	0.00000	1.00000
20	45	0	12	0.00000	.	.

The rest

21	25	1	0	1.00000	.	.
22	25	1	1	0.99953	0.99727	1.00000
23	25	1	2	0.99679	0.98545	1.00000
24	25	1	4	0.99380	0.96712	1.00000
25	25	1	5	0.97393	0.92908	1.00000
26	25	1	7	0.01220	0.00080	0.18538
27	25	1	8	0.00000	0.00000	1.00000
28	25	1	10	0.00000	0.00000	1.00000
29	25	1	11	0.00000	0.00000	1.00000
30	25	1	12	0.00000	.	.
31	45	1	0	1.00000	.	.
32	45	1	1	1.00000	1.00000	1.00000
33	45	1	2	1.00000	0.99998	1.00000
34	45	1	4	0.99999	0.99995	1.00000
35	45	1	5	0.99998	0.99990	1.00000
36	45	1	7	0.99621	0.98945	1.00000
37	45	1	8	0.95800	0.88352	1.00000
38	45	1	10	0.84737	0.67929	1.00000
39	45	1	11	0.38657	0.09793	1.00000
40	45	1	12	0.00000	.	.

Making a plot

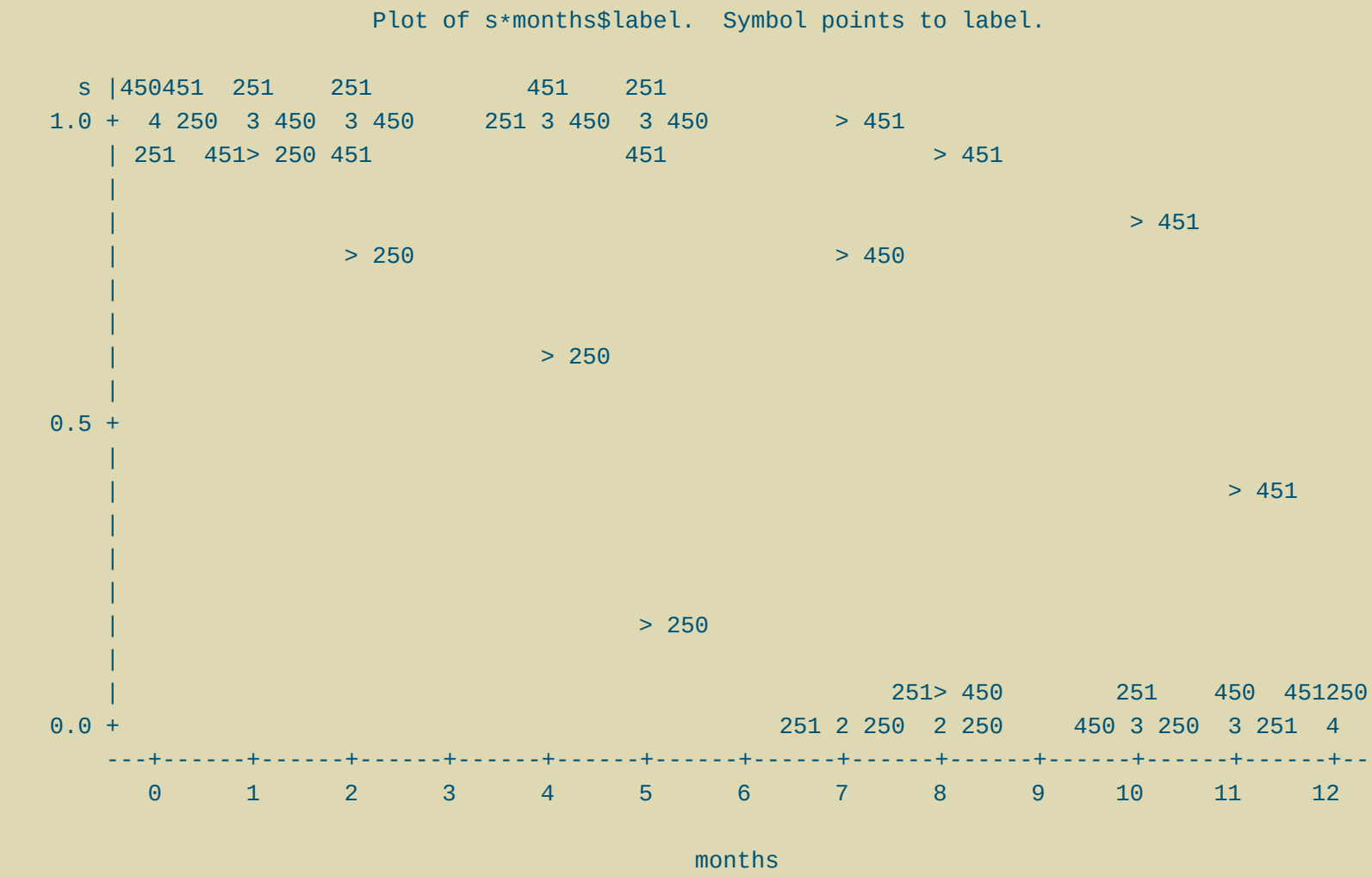
- Start from output data set “fred”.
- Create a new data set with everything in “fred” plus a new variable “label” which will actually be plotted
- Plot the predicted survival probs against “months” marking each point using the labels.

```
data y;
  set fred;
  label=cat(age,treatment);

proc plot;
  plot s*months $ label;
```

- Each plotted point will be labelled with something like 450 (45 year old woman in control group).

The plot



NOTE: 3 label characters hidden.

Discussion

- The confidence intervals are very wide (a lack of data).
- “baseline” doesn’t require any specification of lifetimes, and output gives more detail.
- Look at combinations of treatment and age, see where predicted survival probs “drop off”:
  - age 25, control: after 4 months
  - age 25, treatment: after 5 months
  - age 45, control: after 7 months
  - age 45, treatment: after 10 months
- Indicates definite effect of age, possible effect of treatment.
- Suggests would be worthwhile to do another experiment with more women to get more definite conclusions.