Breast Cancer Study Group

January 14, 2022

Introduction

The assignment consists in a practical analysis of a survival data set. Please prepare a concise and self-contained report , where all answers to the exercises are provided. Also collect all syntax used to analyse the data. Please send the report and the syntax file with comments to Eveline van Beekhuizen $E.S.van_Beekhuizen@lumc.nl.$

Breast cancer data

The data are from the German Breast Cancer Study Group 2 trial. It is a randomized clinical trial on 686 women, comparing hormonal therapy yes/no. Description of the data:

- `horTh`: hormonal therapy, a factor at two levels no and yes
- `age`: of the patients in years.
- `menostat`: menopausal status, a factor at two levels pre (premenopausal) and post (postmenopausal).
- `tsize`: tumor size (in mm).
- `tgrade`: tumor grade, a ordered factor at levels $I < II < III$.
- `pnodes`: number of positive nodes.
- `progrec`: progesterone receptor (in fmol).
- `estrec`: estrogen receptor (in fmol).
- `time`: recurrence free survival time (in days).
- `cens`: censoring indicator (0: censored, 1: event)

Patients data are in the file `GBSG2.txt`. VARIABLE LABELS and VALUE LABELS are provided in the SPSS file `GBSG2.sps`. The table shows the first 6 patients in the data.

id	horTh	age	menostat	tsize	tgrade	pnodes	progrec	estrec	time	cens
1	no	70	Post	21	II	3	48	66	1814	1
2	yes	56	Post	12	II	7	61	77	2018	1
3	yes	58	Post	35	II	9	52	271	712	1
4	yes	59	Post	17	II	4	60	29	1807	1
5	no	73	Post	35	II	1	26	65	772	1
6	no	32	Pre	57	III	24	0	13	448	1

Exercise 1 — Calculate median follow-up of the trial by using the Kaplan-Meier with status indicator reversed. Show the reverse Kaplan-Meier plot. What is its interpretation?

Exercise 2 — Make Kaplan-Meier plots for the two treatment groups. Compare survival with the log-rank test, and calculate a hazard ratio with a 95% confidence interval. What is your conclusion?

Exercise 3 — Perform a multivariate Cox regression using forward selection, excluding the randomized treatment, but including `age`, `menostat`, `tsize`, `tgrade`, `pnodes`, `progrec`, and `estrec` as prognostic variables. If groups defined by certain covariate levels are too small (say smaller than 15) it is wise to merge them with a similar group. Report which variables are added and at what step. Finally, add randomized treatment to this model. Report the final model (hazard ratios, 95% confidence intervals). Has the treatment effect changed with respect to the univariate model? Motivate your answer.

Exercise 4 — The resulting multivariate model (excluding randomized treatment) can be made into a risk score, by adding the regression coefficients multiplied by the covariate values. Calculate the risk score for each patient, and make a histogram.

Exercise 5 — Divide the patients into three groups of roughly equal size with respect to this risk score, call them ”low risk”, ”medium risk” and ”high risk” groups. Make Kaplan-Meier plots for the three risk groups, compare the groups with the log-rank test and report hazard ratios of medium and high risk groups with respect to low risk groups, along with 95% confidence intervals.