Exercises II

Spend a few hours with the dataset *melanoma.dta* . This is from a Danish follow-up of patients treated for malignant melanoma. They were all given the same treatment, and interest is in prognostic factors for survival. Variables are:

Sex: 0 female, 1 male
Survtime: observation time in days
Status: 1 dead from melanoma, 0 censored or dead from other causes
Invasion: depth of invasion, categorical 0,1,2 (Clark Level)
Ici: (please ignore this, I do not know for sure what it means)
Ecells: 1 if tumor contains epithelial cells, 0 otherwise
Ulcerat: 1 if ulceration, 0 otherwise
Thick: tumor thickness in 0.01 cm
Age: age of patient at start of follow-up

The *stset* can be done simply, as *stset survtime, fail(status==1) id(id).* (You will need to create *id* by *gen id = _n*). Create categorical variables for *thick*, cutting at *0, 2, 5, 50* and *age*, cutting at *0, 35, 55, 65, 110*. Before fitting any models, plot the smoothed hazard functions by each of the categorical variables (separately): *sex, invasion, ecells, ulcerat, thickcat, agecat.* You will find rather strong and complicated dependencies on these variables, and a point of the analysis is to avoid over-interpretation of the data. Fit a Cox regression in simplest form (no time-dependent covariables) using all of these categorical variables simultaneously.

Pay some special attention to using only the variables *thickcat* and *invasion*, which presumably are closely related. By looking at the reduction in deviance from fitting each of these, adjusted for the other, consider whether either of these seems to be carrying most of the information from both.

To gain some perspective on the precision with which one can estimate the hazard functions, make a plot of the smoothed hazard without regard to any categorical variables, including pointwise confidence limits for the hazard. Now, using *stsplit,* consider using smoother models in terms of *survtime*. Do the *stsplit* by cutting at 250-day categories. (In this case, in contrast to what happened for the *diabetes* data with more complex *stset*, the variable *survtime* is properly updated for the new records as exit time, and is the same as the variable *_t*) . You do not need to use the *collapse* command.

Use Poisson regression with the new categorical variable for survival time, ignoring other covariables. Create new variables *log(survtime/2000)* and the square of this, and using Poisson regression fit a second-degree polynomial to the hazard function using these (ignoring the other covariables). Create another new covariable that is *log(survtime/2000)* for *survtime <= 2000* and 0 otherwise (that is, truncating the previous covariable). Fit a second-degree polynomial in this new variable. Then fit a model where the hazard function is constant in survival time. Consider the loglikelihoods for the four fits you have made — you will find that there are not large differences in these. Graph (on the same plot) the fitted hazard functions for of the three smooth these fits (enter the fitted parameter values from the keyboard for computing these fitted hazards).

Finally, using Poisson regression and time-constant hazard, fit all the categorical variables used initially for the Cox regression, and compare the parameter estimates and standard errors. You will find that they are virtually the same in the two approaches, illustrating what is called the “efficiency” of Cox regression — it provides estimates of the *RR* essentially as good as those based on a parametric model.

Now return to analyzing the prognostic variables, using Cox regression with no time-dependent covariables in view of what you have just learned. In particular, see if you can reduce the categorical variables *agecat* and *thickcat* to use fewer levels. Generally speaking, try to obtain as simple a model as you can, which maintains most of the prognostic information.