**Survival Analysis – Winter 2019**

**Assignment 2**

**Description of MELANOMA data set (for questions 1 and 2):**

All patients who visited a clinic in 2013 for treatment of melanoma (skin cancer) were followed up from their biopsy date. Their last known vital status was recorded along with the vital status date.

**Variables:**

id: Chart ID

biopsydate: Biopsy date

vstatus: Vital status - Alive, Dead

vstatusdate: Vital status date

/\* covariates: categorization of biopsy results (some data may be missing) \*/

Clarklevel: Clark level (tumour staging)

I=Involves epidermis only

II=Spread somewhat to upper dermis

III=Spread to most of upper dermis

IV=Spread to lower dermis

V=Spread to subcutaneous fat

Ulceration: Presence of skin ulcers

0=No

1=Yes

Thickness: Tumour thickness (mm)

**Question 1:**

* 1. Calculate the outcome time to death in years and check that all values are valid. Patients alive at their vital status date are censored. What proportion of these patients died (using the vital status indictor)? [Note: this proportion is not adjusted for length of follow up.]
  2. Explore the baseline covariates Clark level, ulceration and thickness. Describe the patients’ clinical characteristics using descriptive summaries of each of these covariates. What are the relationships amongst the covariates?
  3. Develop exponential, Weibull, log-logistic, log-normal models for the outcome time to death unadjusted for covariates (i.e. overall survival). Also, develop the gamma model to help decide between the four previous models [see lecture slides 80 and 84] as well as by using the AIC or likelihood ratio test.
  4. Select one of the four parametric models and explain why this is the best model for the data set.
  5. Develop a multivariate model using the chosen parametric model and some or all of the covariates described above. Some patients have missing data. Explain the choices you made to handle the missing data. Explain any transformation of the covariate data.
  6. Assess the goodness of fit of the final model.
  7. Interpret the results of the final model and identify which groups of patients have the best and worst survival. Include a description of the hazard function.

**Question 2:**

1. Fit a log-logistic survival model with the covariate ulceration (excluding the missing cases). (This is not necessarily the correct model for question 1).
2. Produce two survival plots from this log-logistic model (i.e. one for ulceration=yes and one for ulceration=no).
3. What is the estimated time ratio and odds ratio for survival and their 95% confidence intervals using the model parameter estimates (for ulceration compared to no ulceration)?
4. Demonstrate the time ratio using the estimated median survival for each group. Are the estimated medians observed time points in the data?
5. Demonstrate the odds ratio using the estimated proportion surviving 3 years or more.
6. How would you describe the time ratio and odds ratio to a member of the study team who does not have a background in statistics?