**Survival Analysis – Winter 2017**

**Assignment 2**

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

All patients who visited a clinic in 2011 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 time to death in years. Patients alive at the end of follow up are censored. What proportion of these patients died?
  2. Explore the baseline covariates Clark level, ulceration and thickness. Describe the patients’ clinical characteristics using each of these covariates. What are the relationships amongst the covariates?
  3. Develop exponential, Weibull, log-logistic, log-normal and gamma survival models for the outcome time to death unadjusted for covariates (i.e. overall survival).
  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 why you decided to either keep or delete these patients in the analysis. 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 including a description of the hazard function.

**Question 2:**

1. Fit a log-normal 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- normal model (i.e. one for ulceration=yes and one for ulceration=no).
3. What is the estimated median survival for each group and 95% confidence intervals? Are these estimated medians observed time points in the data?
4. What is the estimated time ratio for survival and 95% confidence interval?

**Question 3:**

In multiple myeloma disease the level of hemoglobin in the blood at diagnosis is related to prognosis. In an observational study, one group has low hemoglobin (<12 grams per deciliter) and the other group has high hemoglobin (≥12 grams per deciliter).

Follow up begins at start of treatment. The outcome is time to death and patients who are alive at the end of the study are censored.

There was some concern that the hazard ratio varied across time and log-logistic survival model was fit to the multiple myeloma data. The estimate of the time ratio (low versus high) is 0.48 with a 95% confidence interval of (0.26, 1.14) and a p-value=0.02.

a) Suppose the median survival is 2.3 years for the high hemoglobin group. What is the median survival for the low hemoglobin group?

b) If the scale parameter in the log-logistic model is estimated to be 0.67, what is the estimated constant odds ratio of survival (low versus high)?

c) Compare and interpret the odds ratio of survival and the time ratio for a member of the study team that is not a statistician.