Homework 1 – Applied survival analysis

Consider the dataset *dataHIV.csv* (available on the eclass of the course) which includes survival data from a cohort study of people with HIV. A description of the variables is given below:

- PATIENT: code identifying patient
- mode: Risk group: MSM (men having sex with men), MSW (heterosexuals), PWID (people who inject drugs)
- death: failure indicator (1=death, 0 = censoring)
- time: Time from HIV diagnosis to death or censoring (years).
- CD4: Absolute number of CD4 cells

Please answer the following topics presenting output from ***R***, as well. ***R*** code should be available in an appendix at the end of your document.

1. (**20 Points**) Plot the KM estimates for the risk groups. Which group seems to be doing better in terms of survival? Please use a self-explanatory figure (with an appropriate legend).

2. (**20 Points**) Perform both the Logrank and the Wilcoxon test to compare the survival functions of the risk groups. What do you conclude?

3. (**20 Points**) Check whether the hazards are proportional between the risk groups using the estimated cumulative hazard functions.

4. (**10 Points**) Fit a Cox regression model for the time to death including the risk group and CD4 counts. Interpret the hazard ratios.

5. (**10 Points**) Fit a Cox model assuming also an interaction between the risk group and CD4 counts. This can be easily performed by `Surv(time,death) ~ mode*CD4` in the model's formula, with `mode` being a `factor`. Interpret the effect of CD4 counts for the three groups.

6. (**20 Points**) Ignoring interaction, evaluate the correct functional form of CD4 in the presence of risk group. Is it better on the original scale, log10(x+1) scale, or on the square root scale? Other ideas?