MSCA31010: Linear & Non-Linear Models

Winter 2021 Assignment 4

# Questions 1 and 2

Krall, Uthoff, and Harley (1975) analyzed data from a study on multiple myeloma in which researchers treated sixty-five patients with alkylating agents. Of those patients, forty-eight died during the study, and seventeen survived.

The data set is in the myeloma.csv. The variable Time represents the survival time in months from diagnosis. The variable VStatus consists of two values, 0 and 1, indicating whether the patient was alive or dead, respectively, at the end of the study. If the value of VStatus is 1, the patient died during the study. If the value of VStatus is 0, the patient was still alive at the end of the study and the corresponding value of Time is censored.

The following nine variables thought to be related to survival are

1. LogBUN: logarithm of blood urea nitrogen at diagnosis,
2. HGB: hemoglobin at diagnosis,
3. Platelet: platelets at diagnosis: 0=abnormal, 1=normal,
4. Age: age at diagnosis, in years,
5. LogWBC: logarithm of the number of white blood cells at diagnosis,
6. Frac: fractures at diagnosis: 0=none, 1=present,
7. LogPBM: logarithm of the percentage of plasma cells in bone marrow,
8. Protein: proteinuria at diagnosis, and
9. SCalc: serum calcium at diagnosis.

Our interest lies in identifying important prognostic factors from these nine explanatory variables.

Reference: John M. Krall, Vincent A. Uthoff, and John B. Harley (1975). “A Step-Up Procedure for Selecting Variables Associated with Survival.” *Biometrics*, volume 31, number 1, pages 49 – 57.

# Question 1 (50 points)

1. (30 points). Please use the Kaplan-Meier Product Limit Estimator to create the life table. The life table should include these columns: *Survival Time*, *Number Remaining*, *Number of Deaths*, *Number of Censored*, *Number at Risk*, *Probability of Survival*, *Probability of Failure*, *Standard Error for Probability of Survival*, *Lower 95% Confidence Limit for Probability of Survival*, *Upper 95% Confidence Limit for Probability of Survival*, and *Cumulative Hazard*. Please limit the number of decimal places to four.
2. (10 points). Please generate the Survival Function graph. Since we measure the Time variable in the number of months, we will specify the x-axis ticks from 0 in an increment of 12. Besides plotting the Survival Function versus Time, please also add the 95% Confidence Band. Please consider using the fill\_between() function in matplotlib to generate the Confidence Band as a band around the Survival Function. To receive the full credits, you must label the chart elements properly.
3. (10 points). Please generate the Cumulative Hazard Function graph. To receive the full credits, you must label the chart elements properly.

# Question 2 (50 points)

We will train a Proportional Hazard model to identify important prognostic factors from the nine explanatory variables. We will consider all explanatory variables as interval predictors. We will use the Backward Selection method to exclude non-significant predictors.

We first include all nine explanatory variables in the model. After we have trained a Proportional Hazard model, we will retrieve the summary object to obtain the test significance values (i.e., the p-value) of the explanatory variables. Next, we will look among the explanatory variables whose p-values are more than 0.15. Then, remove the explanatory variable, if any, that has the highest p-value from the model. We will repeat the Backward steps until there are no more explanatory variables that we can remove.

1. (30 points). Please provide a Step Summary table. The table will show the explanatory variable removed from each step and the variable’s test significance value.
2. (10 points). Please provide a Parameter Estimates table that shows the explanatory variables included in the final model. Besides, show these statistics: parameter estimates, standard errors, p-values, Hazard Ratios, and 95% confidence interval for the Hazard Ratios.
3. (10 points). Please plot the estimated baseline cumulative hazard from the Proportional Hazard model versus the observed times. We will also overlay the cumulative hazard function from the Kaplan-Meier estimator for comparison. To receive the full credits, you must label the chart elements properly.