HW1

Business Data Analytics

Due February 12, 2024 37 points

The data consists of measurements made on patients with malignant melanoma. Each patient had their tumor removed by surgery during the period 1962 to 1977. The surgery consisted of complete removal of the tumor together with about 2.5cm of the surrounding skin. Among the measurements taken were the thickness of the tumor and whether it was ulcerated or not. These are thought to be important prognostic variables such that patients with a thick and/or ulcerated tumor have an increased chance of death from melanoma. Patients were followed until the end.

This data frame contains the following columns:

`Time_month:` Survival time in months since the operation, possibly censored.

`Status:` The patient status at the end of the study. 1 indicates that they had died from melanoma, 2 indicates that they were still alive and 3 indicates that they had died from causes unrelated to their melanoma.

`Gender:` The patients' Gender; 1=male, 0=female.

`Age:` Age in years at the time of the operation.

`Year:` Year of operation.

`Thickness:` tumor thickness in mm.

`Ulcer:` Indicator of ulceration; 1=present, 0=absent.

(i) Plot the Kaplan Meier curve for survival. Do the same for male vs female [5]

(ii) Estimate a Cox regression to estimate the impact of the covariates. Write the equation and then estimate the model. How will you interpret these estimates? [10]

(iii) Can you provide some rationale for testing the proportionality assumption? How would you do it in Python? Now provide the results of the test and comment on what you find. [7]

(iv) In this model, you assume that the shape of the hazard is the same and the tumor thickness shifts it proportionally. However, it is possible that the shape of the hazard is different for different tumor thickness. You plan to stratify based on thickness. Since tumor thickness is a continuous measure, your first create a dummy variable which is 1 when tumor thickness is above median and 0 otherwise.
You now stratify your model based on thickness dummy. Operationalize this in your python code. Provide the output.[10] Why do you think the estimates have changed? [5]