

Project 2: Breast cancer survival prediction

Description

This is a dataset of breast cancer patients from a prospective study. Information including variables 1-14 were collected at the baseline, the column “Survival Months” records the length of following-up, and the column “Status” records the survival status of the patients at the end of their following-up. We are primarily interested in predicting the risk of death based on features 1-14.

1. Age
2. Race
3. Marital Status
4. T Stage: Adjusted AJCC 6th T
5. N Stage: Adjusted AJCC 6th N
6. 6th Stage: Breast Adjusted AJCC 6th Stage
7. Differentiate
8. Grade
9. A Stage: Regional — A neoplasm that has extended; Distant — A neoplasm that has spread to parts of the body remote from
10. Tumor Size: Each indicates exact size in millimeters.
11. Estrogen Status
12. Progesterone Status
13. Regional Node Examined
14. Regional Node Positive
15. Survival Months
16. Status: Dead / Alive

Analytical goal (you may only address some of them but properly addressing more will improve the rate of your report):

1. Using variables 1-14 as the covariates to predict the risk of death.
2. Which factors (features) affect the risk significantly? Are there interacting effects?
3. Evaluate the performance of your model(s). Is your model achieving similar performance between the majority race group “White” and the minority “Black” (or “Black” + “Other”)? If not, could you try to improve the fairness (i.e., reducing the gap of prediction performance between the majority and minority) of your model(s)?

Suggestions and tips:

In the report, you should describe your final model and interpret its parameters in an accurate and useful manner. It is expected that you would first examine the marginal distributions and pairwise relationships between variables (e.g., to check to see whether any nonlinearities are immediately obvious), that you would explore several candidate models, and explain why you selected your model.

Also, you should check for violations of regression model assumptions, influential observations, multicollinearity, etc. This project may involve logistic or survival models not introduced in our class, which gives you a 5 points bonus. It would be great if you could be an active learner and figure out these challenges. In addition, model evaluation and fairness mentioned in the above Point 3 are interesting topics, on which you could try to explore.

It would be helpful to be clear about your motivation for carrying out certain analyses as well as to be clear about interpretations of fitted model parameters. Your report should include a table summarizing parameter estimates associated with your final fitted model, characterizing predictor variables in a way that a reader can clearly understand.

Below you'll find some aspects to be addressed in your report. These are just a few suggestions, but feel free to add your own input/creativity to the analysis:

- Data exploration: descriptive statistics and visualization. You might want to, for instance:
 - o Include a descriptive table with summary statistics for all variables;
 - o Explore the distribution of the outcome and consider potential transformations (if necessary);
 - o See if there are any unusual observations and consider them as potential outliers/influential points.
- In your regression model, be watchful for variables that are highly correlated and be selective in the variables you will include in your analysis.
- Consider selective interactions between variables.
- DO NOT IGNORE MODEL DIAGNOSTICS.