AA502 – Survival Analysis

Homework 1

Matthew Austin

Background

Several hurricanes struck the Gulf Coast, resulting in severe casualty and property damage. One of the major defenses against hurricanes is the coordination and maintenance of pump operations during a critical 48-hour period (four high tides). As employees for the Steering Committee of the Center for Risk Management, your team is tasked with conducting a survival analysis for the pump stations in the area.

Data

The `katrina` dataset contains information collected from 770 pump stations. Pump survival is denoted in the variable `survive`, which is an indicator for whether or not the pump survived the entirety of the storm. There are five potential failure conditions which take the following values in the variable `reason`:

- (0) no failure (this is equivalent to `survive = 1`)
- (1) `flood`: overflow or accumulation of water that submerges the pump station
- (2) `motor`: mechanical failure
- (3) `surge`: onshore gush of water typically associated with levee or structural damage
- (4) `jammed`: accumulation of trash or landslide materials

There are eight factors that may influence the survivability of the pump stations. Not all pumps have each characteristic, but some are available for maintenance or upgrade and denoted as such (where a value of `1` indicates the factor is already present and cannot be upgraded):

- `backup` pump (upgrade available): a redundant system used to protect the station from flooding when the main pump is not operating
- `bridge` crane (upgrade available): allows vertical access to equipment and protecting materials
- `servo` (upgrade available): servomechanism used to provide control of a desired operation through the Supervisory Council and Data Acquisition (SCADA) systems
- `trashrack` cleaner (upgrade available): protects hydraulic structures against the inlet of debris, vegetation, or trash
- `elevation` (maintenance available): elevation of the pump station; may be altered by one foot via maintenance
- `slope`: ravine slope surrounding the pump station
- `age`: difference between the pump's installation date and the current date
- `H1--H48`: pumping status reported by pump stations during a 48-hour emergency period (accuracy of pump status not guaranteed to be error-free)

Assignment

As the Steering Committee for the Center of Risk Management, provide a report and a set of recommendations summarizing the findings of your analysis. The report should include the following information:

- Provide summary statistics for each type of pump station failure. What percentage of pumps survived the hurricane? What percentage of pumps are in each failure? What are the median survival times?
- Plot the survival curves for all pumps and the stratified curves (by `reason`). Discuss anything interesting that you find.
- Do the four types of pump failures have similar survival curves? If not, which ones are statistically different? (To get pairwise tests: in `SAS`, use the `diff = all` option in the `strata` statement of `proc lifetest`; in `R`, use `survminer::pairwise_survdiff()` function, which takes the same arguments as `survdiff()`)
  - **If you are using** `R`, note that using `reason` as your grouping variable includes the no failure group, so obviously you will have a very low p-value for the log-rank test (make sure you understand why this is "obvious"). Thus, when carrying out the test, you'll need to remove the surviving pumps. One way to do this is directly from `survdiff()` with the `subset` argument: `survdiff(..., subset = reason != 0)`. But `pairwise_survdiff()` doesn't have a subset argument, so for both functions, you could just do something like `(..., data = katrina[katrina$reason != 0,])` instead (make sure you understand why this works).
- A coworker proposes binning failures into a water-based (flood/surge) group and a mechanical-based (motor/trash) group. Explain why you agree or disagree with this proposal.
- Create hazard plots and discuss anything interesting that you find.

Questions/topics to know

- What is the outcome in survival analysis?
- What summary statistics can we use for survival data?
- How to compute survival probabilities
- How to read and interpret survival curves and hazard plots
- How to test for differences among survival curves
- Relationships between the different functions used in survival analysis
- What is censoring? What are some of the different types? Why might it happen?