Chapter 1

Introduction to Regression Modeling of Survival Data

Regression Modeling

- Models the relationship between an outcome (dependent variable) and a set of predictor variables.
- Type of outcome determines the model used
  - Continuous → Linear regression
    - Example: white blood cell count
  - Dichotomous → Logistic regression
    - Example: presence or absence of a disease
  - Counts → Poisson regression
    - Example: number of hospitalizations in a year

Time-to-Event Data

- When the outcome is the time until an event, survival analysis methods are needed
- In order to measure survival time, need a beginning point (t = 0) and an ending point (t = T)
  - t = 0 is when the clock starts
  - t = T is when the event occurs and the clock stops.
  - Between t = 0 and t = T, the subject is said to be "at risk" for the event
- Survival time is T, the difference between these two points

Time-to-Event Data

- Survival time usually measured in days, months, or years
  - Use units that can be easily interpreted
    - Example: “Survival time was 120 weeks” vs “Survival time was 2.3 years”
  - Note: be careful when changing units
    - Are there 365 days in a year or 365.25?
    - Is the length of time between Jan 1 and Feb 1 (31 days) the same as between Feb 1 and Mar 1 (28 days)?

Time-to-Event Data

- Time is continuous, so why not use linear regression?

- Example: Your outcome of interest is the time until death following one of two cancer treatments
  - Treat 12 people and count how many weeks they live following treatment
  - Some people may drop out of the study before dying
    - You know those people were alive when they dropped out, but don't know how long they survived.

Time-to-Event Data

- Survival times (+ used to indicate drop-out)
  - Treatment A: 10, 20, 30+, 40, 50
  - Treatment B: 10, 30, 50, 60+, 90+, 100+, 110+

- If we exclude the dropouts, the mean survival time for these two treatments is the same.
- However subjects taking Treatment B appear to be living longer.

- Survival times for five subjects are not completely observed. We need a method that allows us to use the data from those five subjects.

Truncation and Censoring

- In survival analysis, incomplete observations are classified as either censored or truncated.
  - Truncation*:* Timing of the event affects whether a subject can be observed.
  - Censoring: Time of the event is unknown.

- Censored observations are included in the study; truncated observations are not used.

Truncation

- Right truncation (very rare):
  - Only subjects who are no longer at risk for the event are observed.
- Left truncation:
  - Subjects enter the study at a given time and are followed from this delayed entry time. Subjects were at risk before starting the study, but only those subjects still at risk at the start of the study are included.

Censoring

- Right censoring (most common):
  - Know the survival time is greater than some value
  - $T > a$
- Left censoring (rare):
  - Know the survival time is less than some value
  - $T < b$
- Interval censoring:
  - Know survival time is within an interval
  - $a < T < b$