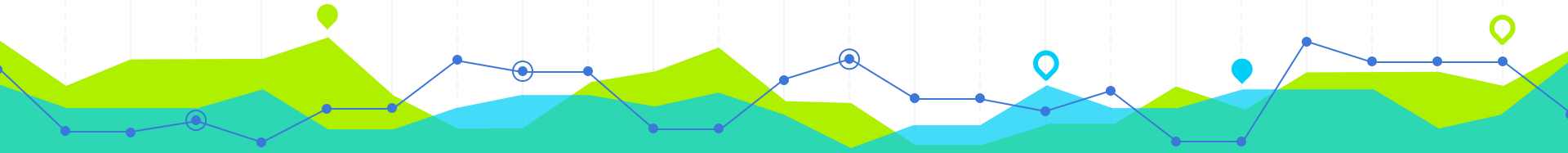


Survival Analysis

Project Introduction

- Explores survival analysis: branch of statistics focused on studying time (from start of observation) until an event occurs
  - *time-to-event data*: Each observation records time until an event occurs or if it occurs at all)

Loosely follows Chapter 7 of Ramzi W. Nahhas' Introduction to Regression Methods for Public Health Using R (2025)

Mentor: Ethan Ancell

Mentees: Alexis and me!

Survival Data/Datasets

Each observation in a survival dataset contains…

- Time-to-event (τ_i) - time until event occurs
- Censoring time (c_i) - time when observation stops (without observing event)
- Observed time/“post-censoring observation” (tilde τ_i) - minimum of τ_i and c_i
- Optional - extra covariates

Introduction to the Natality Dataset

Basic Information

- 2018 United States birth data
- Compiled by National Vital Statistics System with states
- Sample size: Total recorded births in 2018 in the US

Variables

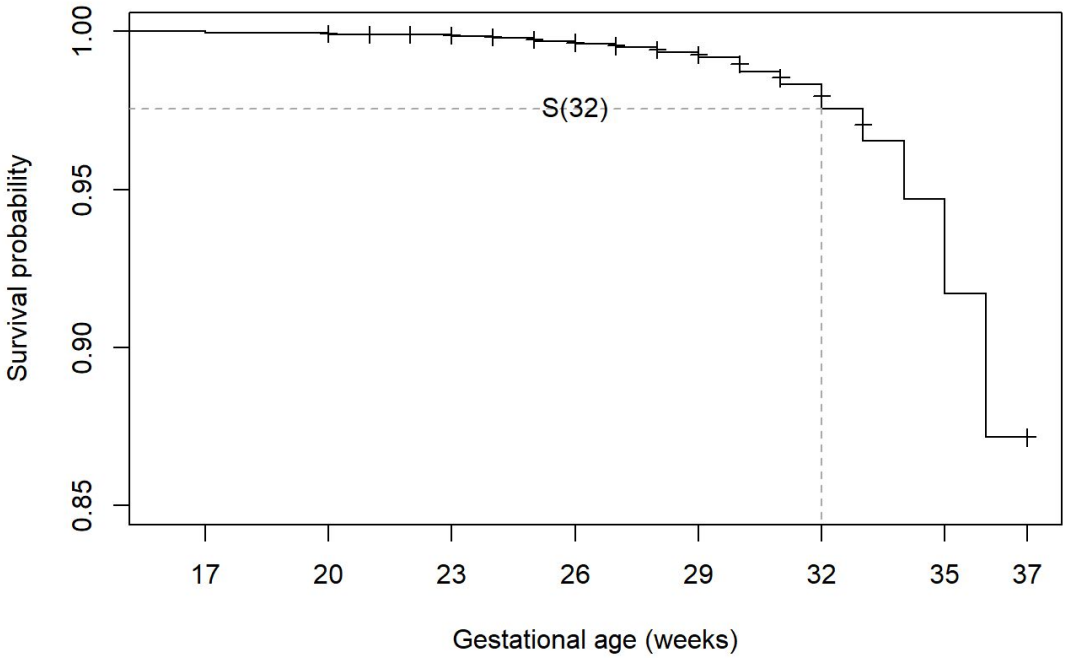
1. gestage37 (numeric) - Gestational age (weeks)
2. preterm01 (binary) - 1 if preterm birth (< 37 weeks), 0 otherwise
3. maternal_age (numeric) - Age of mother
4. smoker (binary) – 1 if mother smoked during pregnancy, 0 otherwise
5. insurance (categorical) – Type of insurance (Private/Public/None)
6. birth_weight (numeric) – Baby birth weight (grams)

Coding Survival Data in R

Two Required Variables

- Survival/event time -> gestage37
- Event indicator -> preterm01 (1 = preterm birth, 0 = censored)

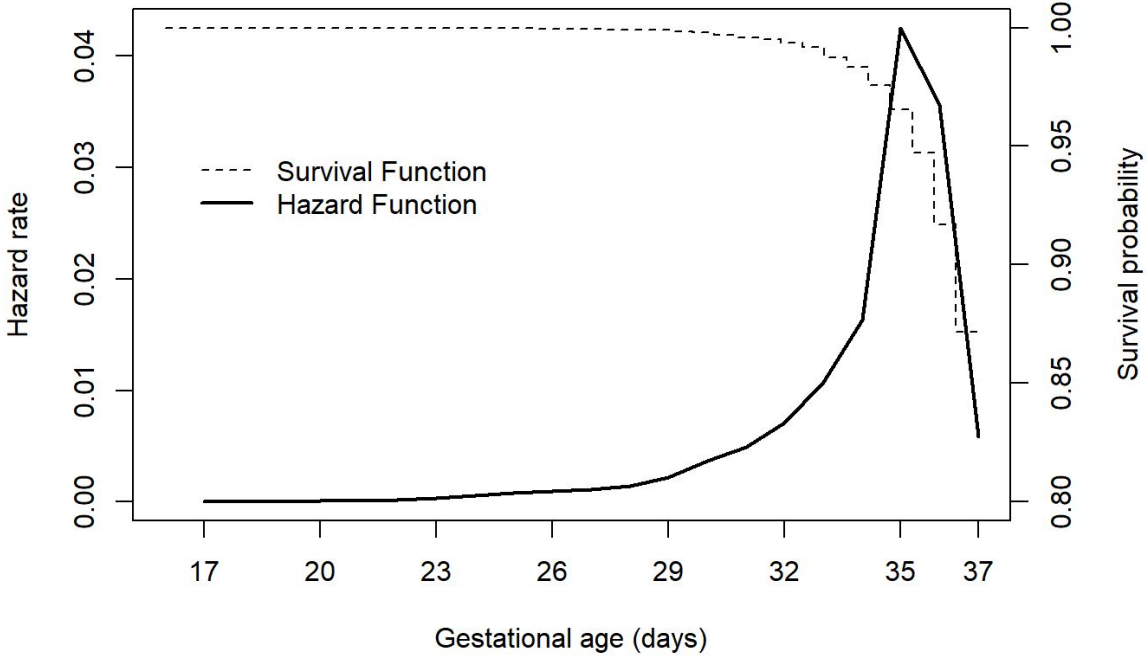
Why? Structure needed for survival models like Kaplan-Meier and Cox regression



S(t) value at t=32 weeks, S(32), is 0.976

- Indicates probability of “survival” (not yet preterm birth) past 32 weeks is 97.6%

Kaplan-Meier S(t) Plot



Can also estimate hazard function

- Notice spike in h(t) corresponds to steeper drop in S(t) estimate (dotted line)

Hazard Function h(t)

THANKS!

Any questions?

Feel free to ask.