Introduction to Survival Analysis

Linyi Xia

(mentor: Kayla Irish)

DRP, Autumn 2024

Our DRP was based on chapter 7 of the online textbook “Introduction to Regression Methods for Public Health Using R” by Ramzi W. Nahhas.

What is Survival Analysis?

A researcher may be interested in understanding **how long it takes for an event to happen**.

- Examples of events: death, disease recurrence, or equipment failure

This is called **time-to-event data**.

Survival analysis is a set of statistical methods designed to analyze time-to-event data.

Time-to-event data we worked with in our DRP

time to preterm birth
(Michalowicz et al. 2006)
Event is preterm birth, time is gestational age

time to heart attack
(Framingham Heart Study)
Event is hospitalized heart attack, time is years from study enrollment to hospitalized heart attack.

time to use heroin
(Carlson et al. 2016)
Event is first use of heroin, time is years from initiation of illicit use of pain pills to first use of heroin

I will focus here today

What Does Survival Data Look Like?

Usually expressed as a vector for each person:

Time-to-event	**Event indicator**
Duration until event occurs	Indicates whether event: • **occurred** (1), or • was **censored** (0)

What is Censoring?

- Occurs when the **exact event time is not observed** for a person.
- There are multiple kinds of censoring, but **we will focus on right censoring**.
- **Right censoring**: the event was not recorded by the study end, and may never occur.

Example of right censoring: if measuring time to preterm birth (birth before 37 weeks), some people in study will not have a preterm birth, so their event is censored.

Preterm birth dataset

gestational week <dbl>	event indicator <dbl>	mom's age <dbl>	race <fctr>	previous preterm birth <fctr>
37	0	35	Hispanic	No
31	1	28	NH White	Yes
37	0	22	NH Black	No
37	0	35	NH White	No
37	0	30	NH White	No

Why Survival Analysis?

- Linear Regression:
  - Cannot handle censored data.
  - Assumes normality and constant variance.
- Logistic Regression:
  - Models binary outcomes but ignores time to event.
- **Survival Analysis**:
  - Accounts for censored data.
  - Describes both probability and timing of events.

Survival Function

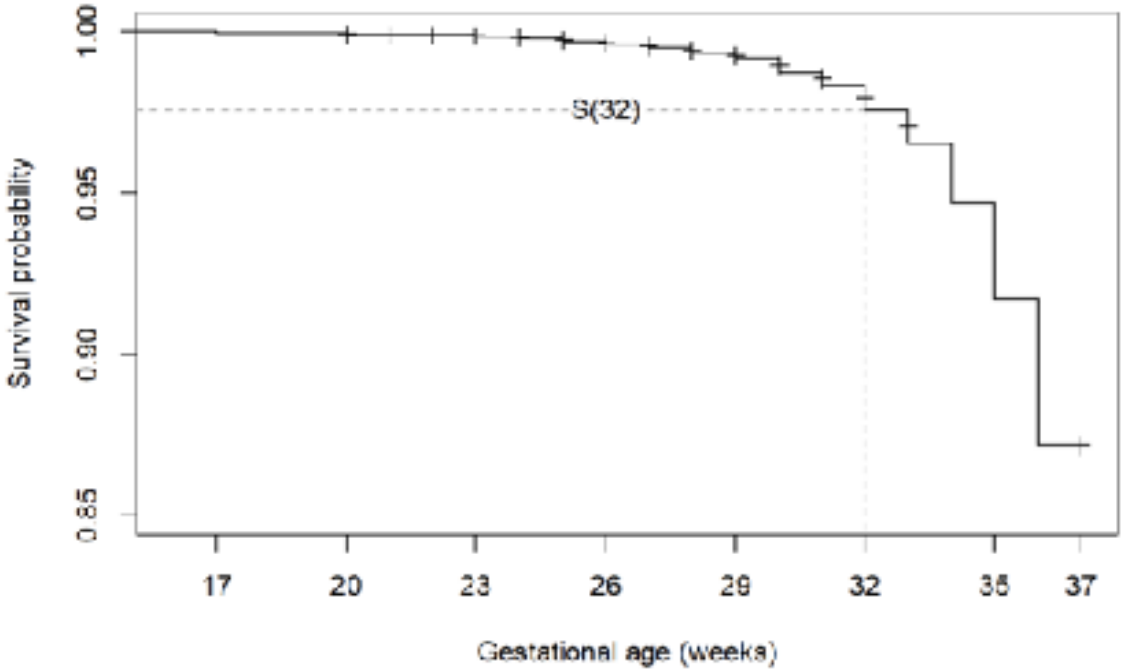
- Survival Function ($S(t)$):
  - Represents the probability of surviving past time t: $S(t) = P(T > t)$.
  - $P(T > t) = 1 - P(T \leqslant t)$.

Kaplan-Meier Estimate

- Kaplan-Meier Estimate:
  - A stepwise function estimating survival function S(t) non-parametrically.

$$
\begin{aligned}
S(t) &= P(T > t) \\
&= P(T > t | T \geq t) P(T \geq t) \\
&= [1 - P(T \leq t | T \geq t)] \, P(T \geq t) \\
&= [1 - P(T = t | T \geq t)] \, P(T \geq t) \\
&= [1 - P(T = t | T \geq t)] \, P(T > t_{\text{prev}}) \\
&= [1 - P(T = t | T \geq t)] \, S(t_{\text{prev}})
\end{aligned}
$$

The KM estimate for the outcome preterm birth



S(32) is 0.976.

This means the estimated probability of "survival" past 32 weeks is 97.6%.

Estimated 2.4% of pregnancies had a preterm birth prior to or at 32 weeks.

Hazard Function

- Definition: Instantaneous rate of event occurrence at time t.
  - Formula:

$$h(t) = lim(\Delta t \to 0)[P(t \leq T < t + \Delta t \,|\, T \geq t)/\Delta t]$$

  - Describes risk of event at a specific time.
  - Complements the survival function.

Cox Regression

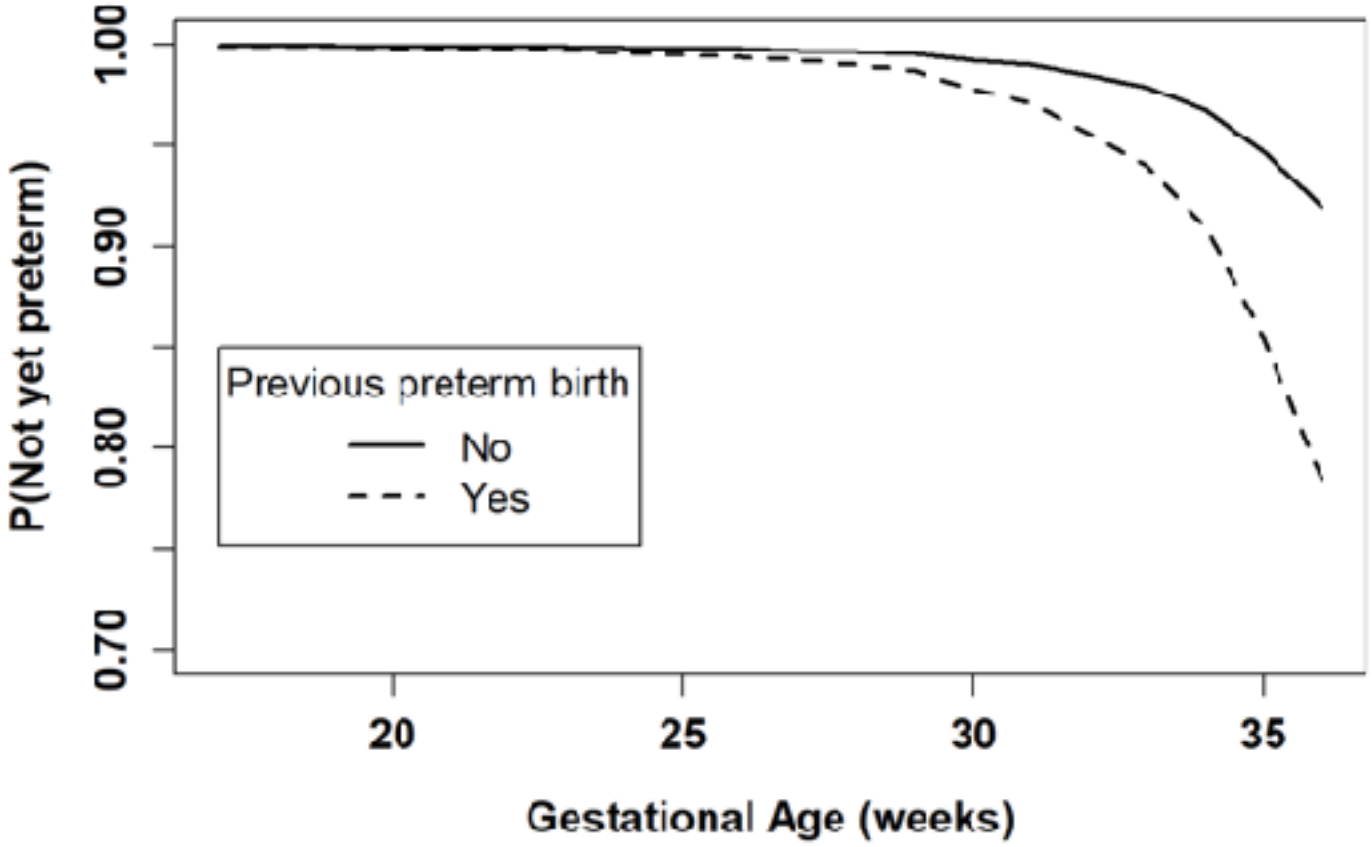
Purpose: Examine relationship between covariates and survival time, accounting for censoring.

Model:

$$h(t \mid X) = h_0(t)exp(\beta 1X1 + \beta 2X2 + \ldots + \beta pXp)$$

where $h_0(t)$ = baseline hazard.

In summary: can get coefficients for covariates, which allows you to interpret the relationship between event risk and a covariate.



Cox Regression can estimate survival functions and account for covariates.

Conclusion

Time-to-event data involves **censoring**, so it needs **special methods** to handle its format.

Kaplan-Meier and **Cox Regression** are two ways to estimate the survival function and understand the relationship between covariates and survival.

Thank you!

Questions?