Biost/Epi 537: Survival Analysis

Discussion Section, Week 5: More KM, Proportional Hazards

February 4, 2025

Using ggplot2 for Kaplan-Meier

- In class, you've seen how to plot KM curves using base R
- Alternative: ggplot2 from the Tidyverse
  - Modify and place legend…
  - Change colors…
  - Display and edit the risk sets…
  - Display and edit censoring marks…
- Some examples in R

Using estimated KM curves in R

- Estimate survival probabilities (with confidence intervals of any significance)
  - Use summary function on a survfit object
- Find arbitrary quantiles (with confidence intervals of any significance)
  - Use summary function on survfit object…
  - Or use quantile function on survfit object

Using ggplot2 for Kaplan-Meier

- ggplot2 is useful when dealing with multiple groups
  - Easy to plot multiple KM estimates of the survival curve on the same graph
  - Easy to plot curves side-by-side or in a grid using “facet”
- Examples in R

Comparing survival probabilities in R

- We saw last time that we can compare survival probabilities at a time t between groups using a Wald statistic:

$$\frac{\hat{T}}{\widehat{SE}(\hat{T})} = \frac{\hat{S}_0(t) - \hat{S}_1(t)}{\sqrt{\widehat{SE}(\hat{S}_0(t))^2 + \widehat{SE}(\hat{S}_1(t))^2}} \approx N(0,1) \text{ under the null.}$$

- So, we can compare $\frac{|\hat{T}|}{\widehat{SE}(\hat{T})}$ to the critical value of $1 - \frac{\alpha}{2}$ of $N(0,1)$ to get a hypothesis test at level α.
- In R?

Log-rank test and variants in R

- As we saw last time, to compare survival curves, we can use the log-rank test.
- In R, this can be done using the survdiff function.
- To do variants of the log-rank test, you could use:
  - comp from survMisc (most direct, but I have had problems using this…)
  - survdiff
  - surv_pvalue from survMiner
- Examples in R

Stratified log-rank test

- In observational studies, confounding is often an issue.
- Confounding is when a third variable causally affects both the exposure/treatment and the outcome (survival).
  - Confounder can't be in the causal pathway between exposure and survival.
- E.g. air pollution study
  - Exposure = pollution level, outcome = pulmonary health, confounder = age
  - Younger people = more likely to live in a more polluted area (affecting exposure), but also less likely to smoke cigarettes (affecting outcome)

Stratified log-rank test

- If we want to compare survival curves between different levels of pollution, the log-rank test wouldn’t account for the effect that confounding by age could have!

- Instead, we use a stratified log-rank test.

  - Looks at the expected vs. observed outcomes within each substrata defined by the confounder – e.g. within young participants and then within old participants.

  - Pools these across different levels of the confounder.

- Example in R

Warning: different null and alternative hypotheses!

- With the log-rank test:
  - H_0 is that $S_0(t) = S_1(t)$ for all t, and H_A is that they differ for at least one t.
- With the stratified log-rank test:
  - H_0 is that, within each level of the confounder, $S_0(t) = S_1(t)$ for all t, whereas H_A is that these differ within at least one level of the confounder.

Regression: Proportional hazards models

- The Kaplan-Meier curve is a great nonparametric estimator, but it doesn't handle extra covariates well.
    - We need to fit different curves for different levels of covariates – this isn't very effective.
- Regression analyses are better equipped for this.
    - Use information across different levels of covariates.
    - Make predictions about covariate values that aren't in the dataset.
        - Be careful about extrapolation!

Proportional hazards assumption

- As always, we need to make assumptions in order to do anything with data.
- A popular assumption in survival analysis is the proportional hazards assumption:
- Two groups satisfy the proportional hazards assumption if their respective hazard functions satisfy $h_1(t) = ch_0(t)$ for all t.
- More generally, given covariates $w = (w_1, \ldots, w_k)$, the assumption is met if $h_0(t \mid w_1, \ldots, w_k) = ch_1(t \mid w_1, \ldots, w_k)$ for all t within all levels of $w_1, \ldots, w_k$.

Is the assumption met in these scenarios?

- Consider a population with brain tumors. The control group isn't treated, while the intervention group receives a risky surgery to remove the tumor. Outcome is time until death.
- Individuals who wear seat belts vs. don't wear seat belts. Outcome is time until death from an automobile accident.
- Individuals who receive vs. don't receive a flu vaccine. Outcome is time until falling sick with the flu.

How can we use proportional hazards?

- Recall that $S(t) = \exp\{-H(t)\}$. Since $H(t) = \int_0^t h(u)du$, the proportional hazards assumption implies that $H_1(t) = \int_0^t h_1(u)du = \int_0^t ch_0(u)du = cH_0(t)$ so that $S_1(t) = \exp\{-cH_0(t)\} = \exp\{-H_0(t)\}^c = S_0(t)^c$.
- This suggests $H_0 : S_1(t) = S_0(t) \iff H_0 : c = 1$.
- If instead we let $\dfrac{h_1(t)}{h_0(t)} = \exp\{\beta\}$, this is equivalent to $H_0 : \beta = 0$.