Final Writeup

This quarter, Nina and I went over materials regarding survival analysis. I started the quarter by getting to know the basic background and concept of survival analysis. Survival analysis is a branch of statistical methods that focus on time-to-event data, or the years, months, weeks, or days from the beginning of follow-up of an individual until an event occurs. In the language of survival analysis, the time until the event occurs is called survival time, and the occurrence of the event is called failure.

When conducting survival analysis, there are many scenarios where the data is not so straightforward. Throughout the rest of the quarter, we gradually delved into more tricky situations and learned about how to deal with each when it occurs.

The most common cases are censoring and truncation. Censoring happens when the exact time of an event of interest is not known, which is usually when the event did not occur throughout the study period or the study lost follow-up with the subject, in which case we call it right censoring. Truncation, on the other hand, occurs when the data only includes a subset of the population. This usually happens when subjects with short survival time cannot make it into the study, causing the study to exclude data from those with short survival time, in which case we call it left truncation. When performing survival analysis, it is important to make sure that censoring is not influencing the survival times or survival outcomes. For example, a censored patient should not withdraw in the middle of the study simply because they felt better over time or felt like the event wouldn't happen. This is called non-informative censoring. In reality, non-informative censoring is difficult to check, as it can neither be easily verified for each subject nor be tested statistically. Therefore, the best researchers can do is to have a reasonable study design so that they are able to assume non-informative censoring.

The survival function and the hazard function are often used to help understand the survival distribution of a certain event. The survival function estimates the probability of a subject at risk surviving until or after a specific time, whereas the hazard function estimates the potential of a subject at risk experiencing failure at a specific time.

Survival functions are often plotted with Kaplan-Meier curves, where the x-axis encodes the time from the start of the study and the y-axis encodes the survival rate. In cases when we want to compare the survival distribution between two groups, we can plot two Kaplan-Meier curves into a single graph to visualize the difference. To study the difference more rigorously, the log-rank test is often used in this case. The log-rank test compares the observed and expected number of events in each group and returns a p-value for the null hypothesis that the two groups share the same survival distribution.

However, there are also some limitations with Kaplan-Meier curves and the log rank test. For example, when the number of factors we are interested in is more than two, or when the factors have multiple levels such as age, it would be messy to draw all the stratified Kaplan-Meier curves into one single plot. Moreover, the log rank test can only take on discrete variables, and it is unable to provide an association parameter for each variable.

Cox regression is a great solution for these limitations. Cox regression will return the hazard ratio of each parameter, which is the ratio between the hazard of the event for the subgroup and the hazard of the event for the baseline group. In other words, if the hazard ratio for a covariate is 1.5, it suggests that the instantaneous risk of the event is 50% higher in the group with the higher unit value of the covariate compared to the group with the lower unit value of the covariate.

As Cox regression only returns one hazard ratio for each covariate, it is essentially making the Proportional Hazards Assumption, which assumes that the hazard ratios between the subgroups and the baseline group stay the same throughout the study. For this reason, it is important to use diagnostic methods to check if this assumption holds when a Cox regression is performed, using metrics like the Schoenfeld residual. If the assumption turns out to be violated, one way to deal with the violation is to stratify the analysis by the variable that violates the assumption, or in other words, perform independent analysis on each level of the variable. Another occasion when the PH assumption fails to hold is when the variable's value changes during the study period, i.e., a time-dependent variable. Using R, we can restructure the data layout for the time-dependent variable, regrouping the period for each individual before and after the time-dependent variable changes its value, thereby allowing us to perform the standard Cox regression.

In summary, we comprehensively covered topics on survival analysis in this quarter. I learned not only the terminologies and methods but also how to perform analysis on time-to-event data on R. This experience was very fruitful and enjoyable, and it matched my expectation at the start of the quarter of learning about survival analysis and training my reading skills.