STATISTIC DEPARTMENT OF UC DAVIS

Project 1, Milestone

Bohao Zou
November 12, 2020

1 IDENTIFY A DATASET

The dataset which I will use comes from the "Survival" package of R. The name of this dataset is "Chemotherapy for Stage B/C colon cancer". Doctors used three different treatments to treat colon cancer. Those three treatments are:

1. ***Obs(ervation)***
2. ***Lev(amisole)***
3. ***Lev(amisole)+5-FU***

In this dataset, there are 929 samples and 9 covariates. Those 9 covariates are:

1. ***sex***: male = 1
2. ***age***: in years
3. ***obstruct***: obstruction of colon by tumour
4. ***perfor***: perforation of colon
5. ***adhere***: adherence to nearby organs
6. ***nodes***: number of lymph nodes with detectable cancer
7. ***differ***: differentiation of tumour
8. ***extent***: Extent of local spread (1=submucosa, 2=muscle, 3=serosa, 4=contiguous structures)

9. ***surg***: time from surgery to registration

There are two records for per person, one for recurrence and one for death. In this analysis, we are only interested in the death event of one patient. There also exists a covariate which name is ***"node4"***, this means more than 4 positive lymph nodes in one patient. "Peter Higgins has pointed out a data inconsistency, revealed by ***nodes, node4***. We don't know which of the two variables is actually correct so have elected not to 'fix' it" in the note of this dataset. So, in this analysis, we only care about the ***nodes*** variable.

2 Define primary questions for analysis

There two primary questions for analysis in this project. The first question is if the hazard rates of different treatments are the same in the statistical significant level of 0.05, if different, which is higher? The second question is which covariates are increase the risk of death and which are decrease the risk?

3 Define the event of interest and survival time

In this dataset, the event of interest is the death of one patient after the treatment. The survival time is the days until event or censoring.

4 Develop the initial data analysis plan

For solving those questions, we should do the following steps:

4.1 FIRST STEP

In out first step, we need to estimate the survival function $S(t)$ of those three different treatments and draw those survival functions in one plot. In our intuition, it is very easy for person to see the distinction between objects in one plot. We can give a first glance about the final result.
There are two ways to deal with it. The one is using a pre-specified distribution and estimate the parameters in that distribution. The other one is Kaplan-Meier or Nelson-Aalen methods. In my consideration, we should use Kaplan-Meier to estimate the survival function. This is because it is hard to assign a appropriate pre-specified distribution and Kaplan-Meier is suitable for estimating the survival function.

4.2 SECOND STEP

In this second step, we can use the hypothesis testing, log-rank test to compare hazard rates of $k = 3$ populations. The reason that why I select log-rank test is that it has optimum power to detect alternatives where the hazard rates in the k populations are proportional to each other.

If the p-value of this log-rank test of $k = 3$ is bigger than or equal with 0.05, we can know that there is no different between the hazard rates of those treatments. If the p-value smaller than 0.05, we can use log-rank test of $k = 2$ to compare those three treatments with each other. It needs 3 times to compare. Because those testing are all independent, we can use Bonferroni correction to control the FWER(Familywise Error Rate) of those hypothesis testing. In the end, we can answer our first primary question if the hazard rates of different treatments are the same and give more specify details that if the there is a different between those three treatments, which two of its are different.

4.3 THREE STEP

For answering the second primary question, we can use Cox-PH model. This is because it is very easy for us to interpret the hazard ratio of each covariates. It can also estimate $\vec{\beta}$ without specifying $h_0(t)$.
By using Cox-PH model, we can know which variable can increase the risk and which can decrease the risk. At the end, we can use Likelihood-ratio test to test those coefficients $\vec{\beta}$ with zero . The meaning of this procedure is to check if those coefficients are truly significant.

5 TAKE A FIRST LOOK AT THE DATASET FOR DATA QUALITY CONTROL.

There are 929 samples in the original samples. There are 41 samples which contain NA value. So, in the end, 888 samples are included in our analysis.

6 FOR SURVIVAL OUTCOME, DEFINE THE FOLLOWING

- Time origin: The time at patient accept corresponding treatment.
- Time scale: days from time origin to death or end of this project.
- Event of interest: Death of the patient.
- Mechanism of censoring and/or truncation: Right censoring.