Jiahui Zeng
Mentor: Antonio Olivas
SPA-DRP
Spring 2022

Estimation For Cancer Screening Model

This quarter, I was selected as one of the participants in the SPA Directed Reading Program where I was paired up with my mentor Antonio Olivas to work on a project. Our goal of the project was to implement a cancer screening program with the greatest benefits, more specifically, a screening program for breast cancer.

During our initial meeting, Antonio gave a brief overview of the problem space which I was half lost due to my lack of knowledge in the field. Therefore, my task for week 1 was to understand the problem by reading papers in the relevant subject. Once I had a better grasp of the background of the problem, we moved on to the statistical concepts, such as exponential distribution, joint continuous distribution, and conditional probability, that were essential in solving our problem.

The readings in the first half of the quarter laid a solid foundation for diving into the actual problem later in the quarter. I began to understand the 3 main questions we wished to answer:

1). When should we start cancer screening? It doesn't give us much information if we screen a younger population because the probability of this subgroup developing breast cancer is nearly 0. Hence, the **age of onset-** cancer appears in the body that is detectable by screening tools- is very important in determining the time of first screening. In our project, we modeled the age of onset with exponential distribution $\mathbf{U \sim Exp(\lambda) + T0}$, where T0 is the minimum age of having breast cancer. This is to assume that women having breast cancer before this age is negligible.

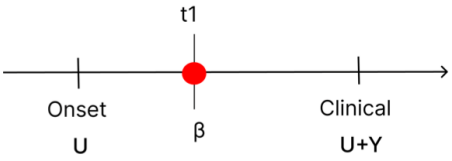
2). How often should we perform screening? If the cancer develops very quickly once it appears in the body, the screening interval should be shorter so that we can catch it before it reaches clinical stage, which is often too late for treatment to cure. In this case, we want to know the average sojourn time- the time from onset to clinical. We modeled the distribution of sourjun time with $\mathbf{Y \sim Exp(\gamma)}$.

3). How accurate is the screening tool? This is called the **sensitivity β**, the probability of detecting cancer when there is cancer, and we are aware that for (1 - β)% of the time, we will miss the cancer cases and place those women in the wrong group. In our project, we assumed that the specificity is 100%, meaning that there's no false positive cases. All the positive cases are truly positive.

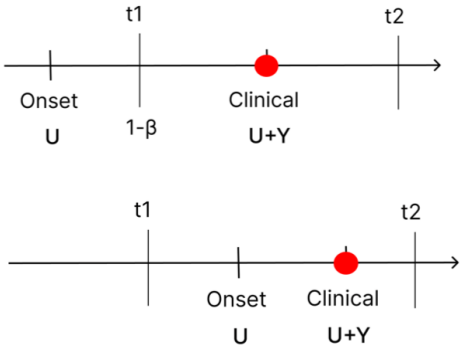
The 3 parameters λ, γ, and β are the key to solving our problem. We would like to obtain these parameters using maximum likelihood estimation. To develop the likelihood function, we first calculated the probability of screen-detected, interval-detected, and not-detected for each screening interval. Each interval is defined as the time from screening at t_i to exactly before the next screening t_{i+1}. Below shows the probability calculation for the first 2 screening intervals. Within each case, there often are multiple scenarios that involve all 3 parameters. The probability of not-detected cases was not drawn below, but it's the probability of 1 minus the other two cases.

$t1 \leq a < t2$

Screen-Detected Case

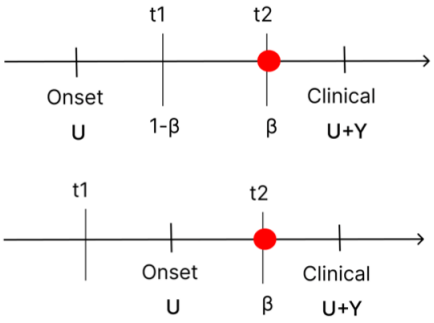


Interval-Detected Case

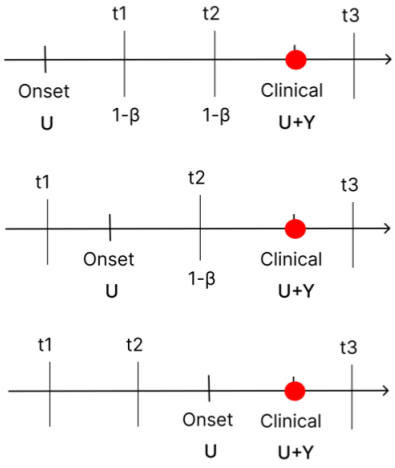


$t2 \leq a < t3$

Screen-Detected Case



Interval-Detected Case



Besides the probabilities, we also fed in real-world data that we gathered from a research paper into our model. Using T0 the minimum age of having cancer at 45 years old, first screening at 50 years old, and the screening interval is one year apart, we estimated the values for the 3 parameters of interest.

```r
data <- rbind(c(142, 15, 19711-142-15), c(66, 10, 17669-66-10))
known <- data.frame(t0 = 45, t1 = 50, t2 = 51, t3 = 52)
known <- cbind(known,
               s1 = data[1, 1],
               r1 = data[1, 2],
               n1 = data[1, 3],
               s2 = data[2, 1],
               r2 = data[2, 2],
               n2 = data[2, 3])
esti <- maxLik(logLik = max_likelihood, start = c(0.001, 0.01, 0.6),
       method = "NR", input = known)

coef(esti)
```

```
[1] 0.00317097 0.19592243 0.72099802
```

With the estimated $\lambda = 0.0031$, we calculated that around 1% of the women in the population have breast cancer at 48.2 years old; $\gamma = 0.1959$ tells us that the mean sojourn time is around 5 years; and $\beta = 0.72$ accounts for the accuracy of the screening tool. Notice that these values were obtained only using 2 screening results. If we were to develop more equations for more screening, we will have a better estimate of the parameters.

If we have more time, we should further evaluate our result to produce a meaningful conclusion. In addition, we should also perform simulation to validate our model to assess how well it performs under a small sample size. A next step challenge is to reevaluate our assumptions about the model and consider alternative methods, such as a non-parametric approach to solve the problem.

Lastly, I can be grateful enough to have Antonio as my mentor for this project, as well as the DRP that provided me with this opportunity to learn more about the biostatistics field. This is a meaningful project that connects what I learn in class with the actual issue we are facing in the real world. In the limited time we had this quarter, we only developed the maximum likelihood

function for 2 screening intervals. I would like to continue on this project by developing the function for all 5 screening intervals. Again, I really appreciate having the opportunity to work on this project, many thanks to my mentor Antonio and the DRP committee!