Mentee: Huayue (Lucia) Zou

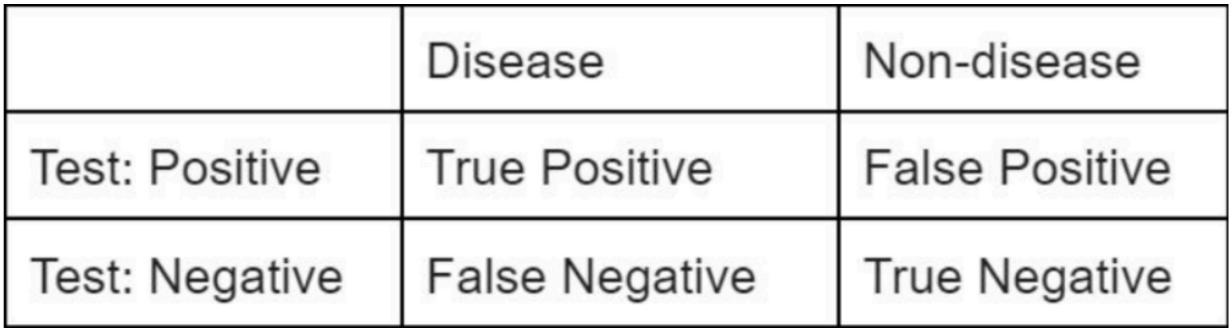
Mentor: Antonio Olivas

STAT499

07 June 2024

Writeup for project: Estimation for cancer screening models using deconvolution

For this whole quarter, I've finished a project with my mentor Mr. Antonio on the topic of "estimation for cancer screening models using deconvolution". In the United States, among every 8 women, 1 of them will develop breast cancer in her daily life. With around 2.5% of the death rate, it's more and more severe in these years. In order to prevent women from getting breast cancer and give their best estimation of physical examination, we need to take a screening test. However, we need to make some selections from the screening test. What does this mean? That is because of the high cost and time spending of screening test, in order to better improve the efficiency of screening and preventions, we need to figure out three questions, and also are the questions that our project is interested in: when should us start screening test, how frequently should us take this screening test again, and how good our test it is. Before we're setting up our functions, we also need to know 3 parameters that we're curious about: gamma (γ), lambda (λ), and beta (β). As for parameter beta, it's the sensitivity of the screening test: it tests the actual probabilities that women have cancer. It's represented as TP/TP+FN, while the corresponding parameters could be learned from figure.1 below.

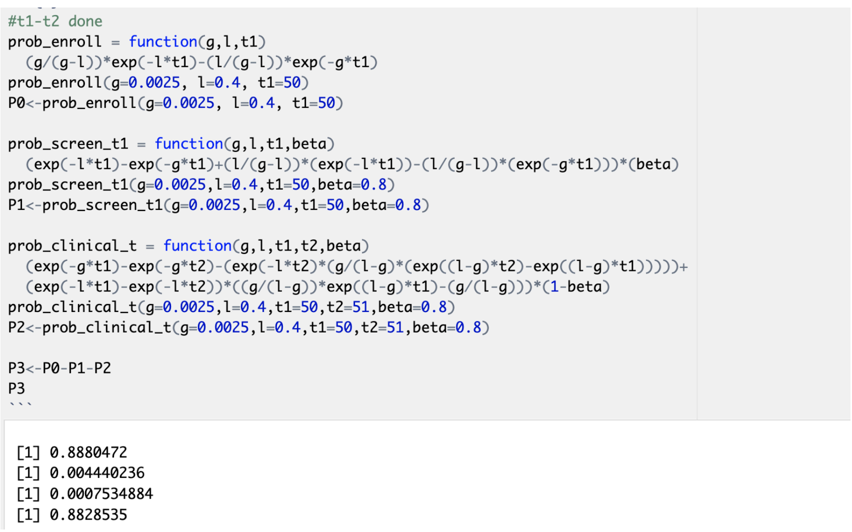


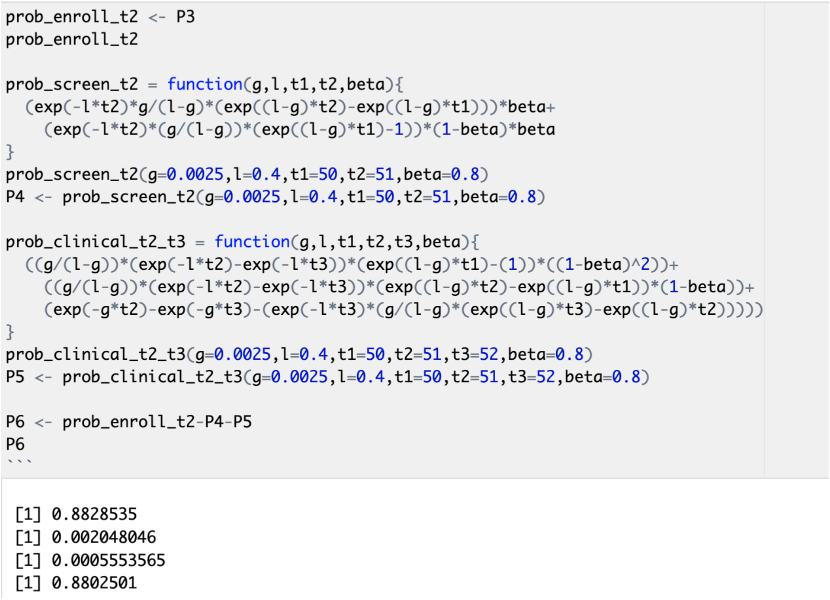
	Disease	Non-disease
Test: Positive	True Positive	False Positive
Test: Negative	False Negative	True Negative

Figure 1.

Then, we’re setting up our exponential distributions’ functions. We suppose the time to onset breast cancer as Y follows exponential distribution with parameter gamma, which can be shown as Y ~ Exp(γ); At the same time, we suppose the Sojourn Time of Breast Cancer as T follows exponential distribution with parameter lambda, which can be shown as T ~ Exp(λ). Then, we’re diving into 5 different time stages to calculate each stage’s probabilities. Before our calculations, in each stage, there are three possible cases: the Subclinical (Screen-detected) case, which suggests that women enroll in the program and have a positive test at a certain time period. The onset time is actually less than the entry time. Secondly is the Clinical (Clinical-detected) case, which suggests that women have symptoms and the onset time is between this time stage. Then the last case is Not-detected, then we’ll move to the next stage as the entry probabilities.

After integrations and calculations, we can get those probabilities from each stage:



```
#t1-t2 done
prob_enroll = function(g,l,t1)
  (g/(g-l))*exp(-l*t1)-(l/(g-l))*exp(-g*t1)
prob_enroll(g=0.0025, l=0.4, t1=50)
P0<-prob_enroll(g=0.0025, l=0.4, t1=50)

prob_screen_t1 = function(g,l,t1,beta)
  (exp(-l*t1)-exp(-g*t1)+(l/(g-l))*(exp(-l*t1))-(l/(g-l))*(exp(-g*t1)))*(beta)
prob_screen_t1(g=0.0025,l=0.4,t1=50,beta=0.8)
P1<-prob_screen_t1(g=0.0025,l=0.4,t1=50,beta=0.8)

prob_clinical_t = function(g,l,t1,t2,beta)
  (exp(-g*t1)-exp(-g*t2)-(exp(-l*t2)*(g/(l-g)*(exp((l-g)*t2)-exp((l-g)*t1)))))+
  (exp(-l*t1)-exp(-l*t2))*((g/(l-g))*exp((l-g)*t1)-(g/(l-g)))*(1-beta)
prob_clinical_t(g=0.0025,l=0.4,t1=50,t2=51,beta=0.8)
P2<-prob_clinical_t(g=0.0025,l=0.4,t1=50,t2=51,beta=0.8)

P3<-P0-P1-P2
P3
```
```

```
[1] 0.8880472
[1] 0.004440236
[1] 0.0007534884
[1] 0.8828535
```



```
prob_enroll_t2 <- P3
prob_enroll_t2

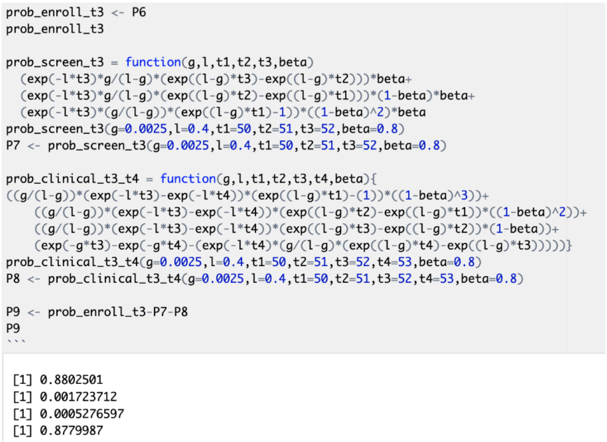
prob_screen_t2 = function(g,l,t1,t2,beta){
  (exp(-l*t2)*g/(l-g)*(exp((l-g)*t2)-exp((l-g)*t1)))*beta+
    (exp(-l*t2)*(g/(l-g))*(exp((l-g)*t1)-1))*(1-beta)*beta
}
prob_screen_t2(g=0.0025,l=0.4,t1=50,t2=51,beta=0.8)
P4 <- prob_screen_t2(g=0.0025,l=0.4,t1=50,t2=51,beta=0.8)

prob_clinical_t2_t3 = function(g,l,t1,t2,t3,beta){
  ((g/(l-g))*(exp(-l*t2)-exp(-l*t3))*(exp((l-g)*t1)-(1))*((1-beta)^2))+
    ((g/(l-g))*(exp(-l*t2)-exp(-l*t3))*(exp((l-g)*t2)-exp((l-g)*t1))*(1-beta))+
    (exp(-g*t2)-exp(-g*t3)-(exp(-l*t3)*(g/(l-g)*(exp((l-g)*t3)-exp((l-g)*t2)))))
}
prob_clinical_t2_t3(g=0.0025,l=0.4,t1=50,t2=51,t3=52,beta=0.8)
P5 <- prob_clinical_t2_t3(g=0.0025,l=0.4,t1=50,t2=51,t3=52,beta=0.8)

P6 <- prob_enroll_t2-P4-P5
P6
```
```

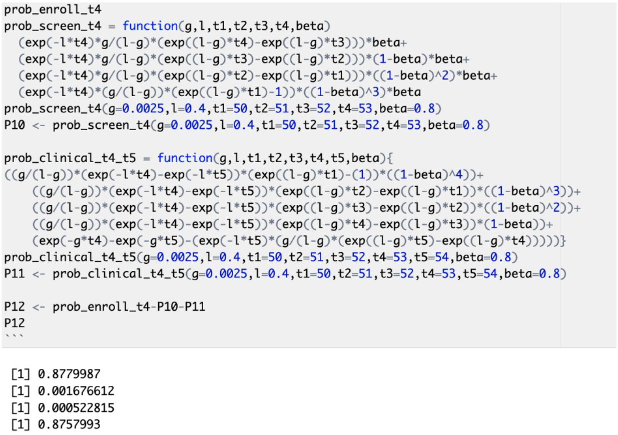
```
[1] 0.8828535
[1] 0.002048046
[1] 0.0005553565
[1] 0.8802501
```

T3 - T4



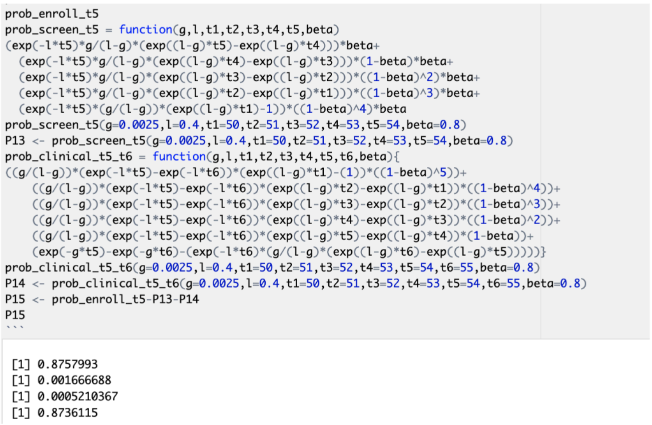
~~~
prob_enroll_t3 <- P6
prob_enroll_t3

prob_screen_t3 = function(g,l,t1,t2,t3,beta)
  (exp(-l*t3)*g/(l-g)*(exp((l-g)*t3)-exp((l-g)*t2)))*beta+
  (exp(-l*t3)*g/(l-g)*(exp((l-g)*t2)-exp((l-g)*t1)))*(1-beta)*beta+
  (exp(-l*t3)*(g/(l-g))*(exp((l-g)*t1)-1))*((1-beta)^2)*beta
prob_screen_t3(g=0.0025,l=0.4,t1=50,t2=51,t3=52,beta=0.8)
P7 <- prob_screen_t3(g=0.0025,l=0.4,t1=50,t2=51,t3=52,beta=0.8)

prob_clinical_t3_t4 = function(g,l,t1,t2,t3,t4,beta){
((g/(l-g))*(exp(-l*t3)-exp(-l*t4))*(exp((l-g)*t1)-(1))*((1-beta)^3))+
    ((g/(l-g))*(exp(-l*t3)-exp(-l*t4))*(exp((l-g)*t2)-exp((l-g)*t1))*((1-beta)^2))+
    ((g/(l-g))*(exp(-l*t3)-exp(-l*t4))*(exp((l-g)*t3)-exp((l-g)*t2))*(1-beta))+
    (exp(-g*t3)-exp(-g*t4)-(exp(-l*t4)*(g/(l-g)*(exp((l-g)*t4)-exp((l-g)*t3))))))}
prob_clinical_t3_t4(g=0.0025,l=0.4,t1=50,t2=51,t3=52,t4=53,beta=0.8)
P8 <- prob_clinical_t3_t4(g=0.0025,l=0.4,t1=50,t2=51,t3=52,t4=53,beta=0.8)

P9 <- prob_enroll_t3-P7-P8
P9
```

 [1] 0.8802501
 [1] 0.001723712
 [1] 0.0005276597
 [1] 0.8779987
~~~

T4 - T5



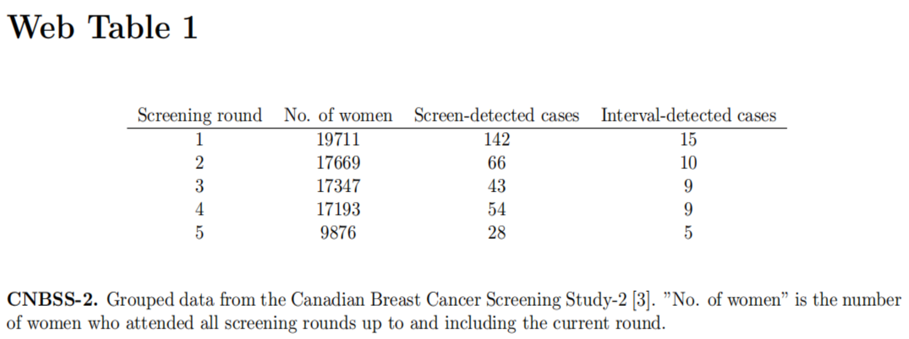
~~~
prob_enroll_t4
prob_screen_t4 = function(g,l,t1,t2,t3,t4,beta)
  (exp(-l*t4)*g/(l-g)*(exp((l-g)*t4)-exp((l-g)*t3)))*beta+
  (exp(-l*t4)*g/(l-g)*(exp((l-g)*t3)-exp((l-g)*t2)))*(1-beta)*beta+
  (exp(-l*t4)*g/(l-g)*(exp((l-g)*t2)-exp((l-g)*t1)))*((1-beta)^2)*beta+
  (exp(-l*t4)*(g/(l-g))*(exp((l-g)*t1)-1))*((1-beta)^3)*beta
prob_screen_t4(g=0.0025,l=0.4,t1=50,t2=51,t3=52,t4=53,beta=0.8)
P10 <- prob_screen_t4(g=0.0025,l=0.4,t1=50,t2=51,t3=52,t4=53,beta=0.8)

prob_clinical_t4_t5 = function(g,l,t1,t2,t3,t4,t5,beta){
((g/(l-g))*(exp(-l*t4)-exp(-l*t5))*(exp((l-g)*t1)-(1))*((1-beta)^4))+
    ((g/(l-g))*(exp(-l*t4)-exp(-l*t5))*(exp((l-g)*t2)-exp((l-g)*t1))*((1-beta)^3))+
    ((g/(l-g))*(exp(-l*t4)-exp(-l*t5))*(exp((l-g)*t3)-exp((l-g)*t2))*((1-beta)^2))+
    ((g/(l-g))*(exp(-l*t4)-exp(-l*t5))*(exp((l-g)*t4)-exp((l-g)*t3))*(1-beta))+
    (exp(-g*t4)-exp(-g*t5)-(exp(-l*t5)*(g/(l-g)*(exp((l-g)*t5)-exp((l-g)*t4))))))}
prob_clinical_t4_t5(g=0.0025,l=0.4,t1=50,t2=51,t3=52,t4=53,t5=54,beta=0.8)
P11 <- prob_clinical_t4_t5(g=0.0025,l=0.4,t1=50,t2=51,t3=52,t4=53,t5=54,beta=0.8)

P12 <- prob_enroll_t4-P10-P11
P12
```

 [1] 0.8779987
 [1] 0.001676612
 [1] 0.000522815
 [1] 0.8757993
~~~

T5 - T6



~~~
prob_enroll_t5
prob_screen_t5 = function(g,l,t1,t2,t3,t4,t5,beta)
(exp(-l*t5)*g/(l-g)*(exp((l-g)*t5)-exp((l-g)*t4)))*beta+
  (exp(-l*t5)*g/(l-g)*(exp((l-g)*t4)-exp((l-g)*t3)))*(1-beta)*beta+
  (exp(-l*t5)*g/(l-g)*(exp((l-g)*t3)-exp((l-g)*t2)))*((1-beta)^2)*beta+
  (exp(-l*t5)*g/(l-g)*(exp((l-g)*t2)-exp((l-g)*t1)))*((1-beta)^3)*beta+
  (exp(-l*t5)*(g/(l-g))*(exp((l-g)*t1)-1))*((1-beta)^4)*beta
prob_screen_t5(g=0.0025,l=0.4,t1=50,t2=51,t3=52,t4=53,t5=54,beta=0.8)
P13 <- prob_screen_t5(g=0.0025,l=0.4,t1=50,t2=51,t3=52,t4=53,t5=54,beta=0.8)
prob_clinical_t5_t6 = function(g,l,t1,t2,t3,t4,t5,t6,beta){
((g/(l-g))*(exp(-l*t5)-exp(-l*t6))*(exp((l-g)*t1)-(1))*((1-beta)^5))+
    ((g/(l-g))*(exp(-l*t5)-exp(-l*t6))*(exp((l-g)*t2)-exp((l-g)*t1))*((1-beta)^4))+
    ((g/(l-g))*(exp(-l*t5)-exp(-l*t6))*(exp((l-g)*t3)-exp((l-g)*t2))*((1-beta)^3))+
    ((g/(l-g))*(exp(-l*t5)-exp(-l*t6))*(exp((l-g)*t4)-exp((l-g)*t3))*((1-beta)^2))+
    ((g/(l-g))*(exp(-l*t5)-exp(-l*t6))*(exp((l-g)*t5)-exp((l-g)*t4))*(1-beta))+
    (exp(-g*t5)-exp(-g*t6)-(exp(-l*t6)*(g/(l-g)*(exp((l-g)*t6)-exp((l-g)*t5))))))}
prob_clinical_t5_t6(g=0.0025,l=0.4,t1=50,t2=51,t3=52,t4=53,t5=54,t6=55,beta=0.8)
P14 <- prob_clinical_t5_t6(g=0.0025,l=0.4,t1=50,t2=51,t3=52,t4=53,t5=54,t6=55,beta=0.8)
P15 <- prob_enroll_t5-P13-P14
P15
```

 [1] 0.8757993
 [1] 0.001666688
 [1] 0.0005210367
 [1] 0.8736115
~~~

With these given probabilities, we could apply them to get our conclusions. Firstly, by using the data from the paper “Identification of the Fraction of Indolent Tumors and Associated Overdiagnosis in Breast Cancer Screening Trials” written by Marc D et al (figure 2).



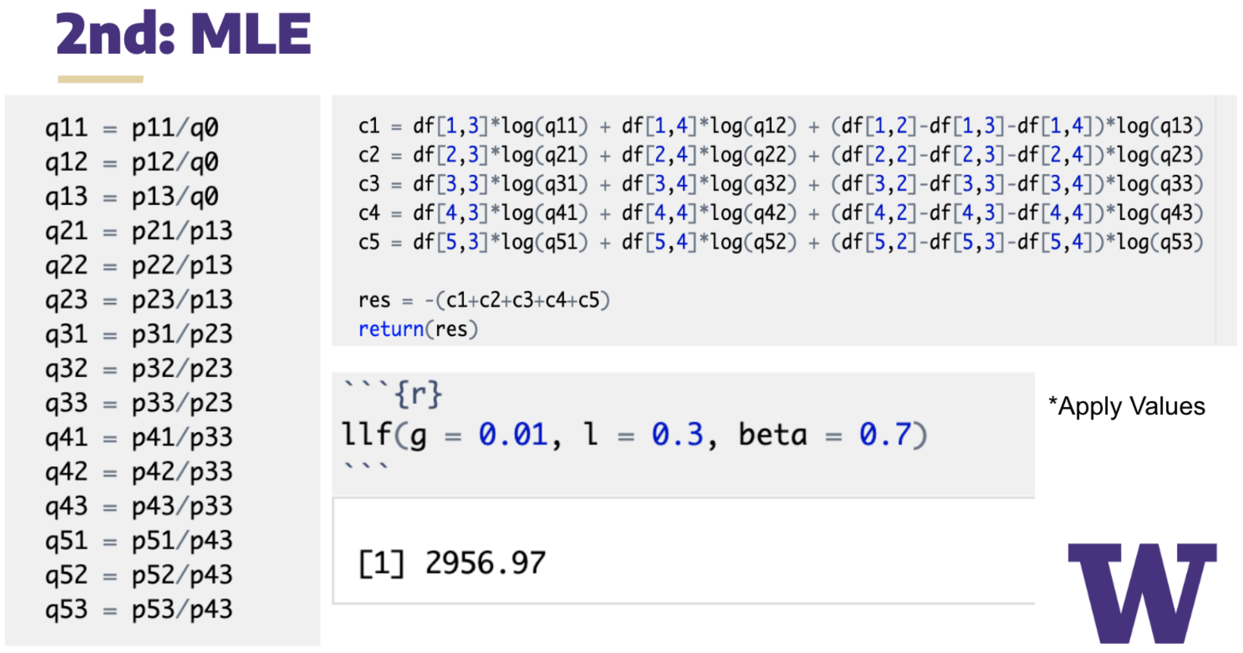
Web Table 1

Screening round	No. of women	Screen-detected cases	Interval-detected cases
1	19711	142	15
2	17669	66	10
3	17347	43	9
4	17193	54	9
5	9876	28	5

CNBSS-2. Grouped data from the Canadian Breast Cancer Screening Study-2 [3]. ”No. of women” is the number of women who attended all screening rounds up to and including the current round.

Figure 2

Then, by using these data to create a data frame, we use Maximum Likelihood Estimation to get our final estimates for 3 parameters. While the process for taking the MLE is to take the LOG of the function first, then calculate probabilities, finally bring it back to res (a method used in R to optimize and find the maximum likelihood estimates).



2nd: MLE

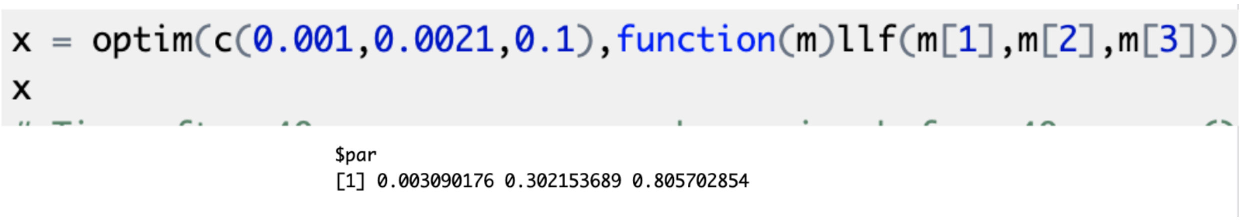
```
q11 = p11/q0
q12 = p12/q0
q13 = p13/q0
q21 = p21/p13
q22 = p22/p13
q23 = p23/p13
q31 = p31/p23
q32 = p32/p23
q33 = p33/p23
q41 = p41/p33
q42 = p42/p33
q43 = p43/p33
q51 = p51/p43
q52 = p52/p43
q53 = p53/p43
```

```
c1 = df[1,3]*log(q11) + df[1,4]*log(q12) + (df[1,2]-df[1,3]-df[1,4])*log(q13)
c2 = df[2,3]*log(q21) + df[2,4]*log(q22) + (df[2,2]-df[2,3]-df[2,4])*log(q23)
c3 = df[3,3]*log(q31) + df[3,4]*log(q32) + (df[3,2]-df[3,3]-df[3,4])*log(q33)
c4 = df[4,3]*log(q41) + df[4,4]*log(q42) + (df[4,2]-df[4,3]-df[4,4])*log(q43)
c5 = df[5,3]*log(q51) + df[5,4]*log(q52) + (df[5,2]-df[5,3]-df[5,4])*log(q53)

res = -(c1+c2+c3+c4+c5)
return(res)
```

~~~
```{r}
llf(g = 0.01, l = 0.3, beta = 0.7)
```
~~~

*Apply Values

```
[1] 2956.97
```



```
x = optim(c(0.001,0.0021,0.1),function(m)llf(m[1],m[2],m[3]))
x
```

```
$par
[1] 0.003090176 0.302153689 0.805702854
```

With our MLE of gamma, lambda, and beta, we then use them to have our conclusions. When should we start the screening test? Since the MLE of gamma is equal to 0.0031 and we suppose choose the prob. of breast cancer that women onset rate is 1.5% & 3%, after transforming CDF of exponential distribution to calculate our estimate years, we get # 1.5% log(0.985)/(-0.003148025) = 4.8 -- start 5 years after 40 years, # 3% log(0.97)/(-0.003148025) = 9.68 -- start 10 years after 40 years (higher prevalence, and could spend less money since screening tests are expensive). Secondly, how frequently should they take the screening test? Since MLE of lambda is approximately 0.3022, our expected years of sojourn time are E[Sojourn time] = 1/0.3022 = 3.3091, which means the average onset time moves to the clinical stage is around 3.3 yrs. Thus, we could suggest taking screening time every 2-3 years. Lastly, how good is our testing? Since the MLE of beta is 0.8057, it represents the sensitivity of the screening test. Thus, we could conclude that for women with breast cancer, there is 81% probability of detecting that actually there is cancer.

Within the future steps, I could dive into calculating the Confidence Interval for catching the true probabilities and test whether the given data is correct or not.

REFERENCES:

Fancher, Tressa. “Breast Cancer Awareness Month.” Breast Cancer Awareness Month, Blogger, 16 Oct. 2017, bplolinenews.blogspot.com/2017/10/breast-cancer-awareness-month.html.

Golden, Richard. “LM101-055: How to Learn Statistical Regularities Using Map and Maximum Likelihood Estimation (Rerun).” Learning Machines 101, 16 Aug. 2016, www.learningmachines101.com/lm101-055-learn-statistical-regularities-using-map-maximum-likelihood-estimation-rerun/

Marc D et al., “Identification of the Fraction of Indolent Tumors and Associated Overdiagnosis in Breast Cancer Screening Trials”