hw8

Timothy

April 19, 2018

Contents

1 Tonsils

1.1 a

Since a two-state mover-stayer model is an aperiodic positive recurrent DTMC, The equilibrium proportion of time spent having infections is equal to steady state vector given by $\lim_{m\to\infty}(\mathbf{\Pi^N}^n$, or more easily calculated: the right eigenvalues of $\mathbf{\Pi^N}$.

We seek the vector $\mathbf{v}$ such that: $\mathbf{\Pi^N}v = v$ This simplifies to this system

of equations:

$$\begin{aligned} v_0 &= v_0\pi_{00} + v_1(1-\pi_{00}) \\ v_0 + v_1 &= 1 \\ (1-v_1) &= (1-v_1)\pi_{00} + v_1(1-\pi_{11}) \end{aligned}$$

$$\begin{aligned} v_1 &= \frac{\pi_{00}-1}{\pi_{00}+\pi_{11}-2} \\ v_0 &= \frac{\pi_{11}-1}{\pi_{00}+\pi_{11}-2} \end{aligned}$$

From this we can see that:

$$\begin{aligned} \pi_1^N &= \frac{\pi_{00}^N - 1}{\pi_{00}^N + \pi_{11}^N - 2} \\ \pi_1^W &= \frac{\pi_{00}^W - 1}{\pi_{00}^W + \pi_{11}^W - 2} \end{aligned}$$

1.2 b

Setting the steady state probabilities equal:

$$\frac{\pi_{00}^N - 1}{\pi_{00}^N + \pi_{11}^N - 2} = \frac{\pi_{00}^W - 1}{\pi_{00}^W + \pi_{11}^W - 2}$$

$$\pi_{00}^N = \frac{\pi_{11}^W + \pi_{00}^W\pi_{11}^N - \pi_{00}^W - \pi_{11}^N}{\pi_{11}^W}$$

Clearly this condition can hold in many situations.

1.3 c

The contribution to the likelihood function of the k$^{\text{th}}$ person will depend on whether or not he has had his tonsils removed, as well as whether or not he is healthty or sick in the (t-1)$^{\text{th}}$ period. Let $J \in \{W, N\}$ denote whether or not the k$^{\text{th}}$ person has had his tonsils removed. Conditioned on whether or not he is healthy or sick in the (t-1)$^{\text{th}}$ period, the contribution to the likelihood function will be a bernouli random variable and is given by:

$$\begin{aligned} P(X_t^J = k | X_{t-1}^J = 0) &= (\pi_{00}^J)^k(1-\pi_{00}^J)^{1-k} \\ P(X_t^J = k | X_{t-1}^J = 1) &= (\pi_{11}^J)^k(1-\pi_{11}^J)^{1-k} \end{aligned}$$

1.4 d

Assume that there is N_W people in the sample, each with T_w data points. The likelihood of the sample is then given by:

$$L(y|\pi^W) = \prod_{n=1}^{N_W}\prod_{t=1}^{T_W} P(X_t^W = y_t|y_{t-1}=0)1_{\{y_{t-1}=0\}} + P(X_t^W = y_t|y_{t-1}=1)1_{\{y_{t-1}=1\}}$$

$$L(y|\pi^W) = \prod_{n=1}^{N_W}\prod_{t=1}^{T_W} \left((\pi_{00}^W)^{y_t}(1-\pi_{00}^W)^{1-y_t}\right)1_{\{y_{t-1}=0\}} + \left((\pi_{11}^W)^{y_t}(1-\pi_{11}^W)^{1-y_t}\right)1_{\{y_{t-1}=1\}}$$

Since at time t, the realization of y_{t-1} is known, instead of being a probability, it is an indicator function stating whether or not he was sick or healthy in the past time period.

We would then take the log of the likelihood function, and calculate its gradient and hessian, then find the values of π_{00}^W and π_{11}^W that maximize the likelihood. However this is quite ungainly because of the sum within the likelihood function. Since y_{t-1} is known, we may partition the data into two sets, S and $\bar{S}$, where S contains the data points where the individual was healthy in the previous time period and $\bar{S}$ is the complement.

$$L(y|\pi^W) = \prod_{y_t\in S}\left((\pi_{00}^W)^{y_t}(1-\pi_{00}^W)^{1-y_t}\right)\prod_{y_t\in\bar{S}}\left((\pi_{11}^W)^{y_t}(1-\pi_{11}^W)^{1-y_t}\right)$$

1.5 e

There are now two indicator functions, one for the state of the last period, and one for the presense of tonsils. So we will need to partition the data into two more sets. Consider $S_0, \bar{S_0}, S_1, \bar{S_1}$ which each contain the data points where S_j is healthy in the previous state, and j denotes if they have tonsils or not.

$$L(y|\pi^W) = \prod_{y_t\in S_0}\left((\pi_{00}^N)^{y_t}(1-\pi_{00}^N)^{1-y_t}\right)\prod_{y_t\in\bar{S_0}}\left((\pi_{11}^N)^{y_t}(1-\pi_{11}^N)^{1-y_t}\right)$$
$$\prod_{y_t\in S_1}\left((\pi_{00}^W)^{y_t}(1-\pi_{00}^W)^{1-y_t}\right)\prod_{y_t\in\bar{S_1}}\left((\pi_{11}^W)^{y_t}(1-\pi_{11}^W)^{1-y_t}\right)$$

1.6 f

We would like to test if there is any difference between the markov chains that define the law of motion in the two state spaces. Therefore we need to test the simultaneous condition that:

$$\pi_{00}^W = \pi_{00}^N$$
$$\pi_{11}^W = \pi_{11}^N$$

Since we already know the likelihood function under both the null and the alternate hypothesis, using a likelihood ratio test seems to be the easiest test. It is well known that: $2(f(\pi^a - f(\pi^0)) \sim \chi^2(2)$ where f is the log likelihood function under each hypothesis.

1.7 g

The analysis conducted so far has not considered the cost of the tonsilectomies conducted. If the benefit from the tonsilectomy is valued at a $100 gain by reducing the probability of a throat infection slightly, but costs $150 to conduct, there is a net loss by having the tonsilectomy conducted. Since we would have found a small change in the reduction of throat infections, and throat infections are relatively uncommon and not particularily devastating, there being a small gain by having this probabilty reduced is very plausible. In this case, the cost of the operation is more than the benefit of receiving the operation, and it is being overperformed.

1.8 h

1.9 i

```
tonsilData <- read.table( 'Tonsils.dat', header=FALSE )

##Firstly, V5 is whether or not they have tonsils
##V1 - D_00
##V2 - D_01
##V3 - D_10
##V4 - D_11

## for the model X will be transformed into a tuple of 4 values:
## X = (start state 0, tonsils; start 1, tonsils; start 0, no tonsils; start 1, no tonsils )
## Y = end at state 1

tonsilModel <- glm( formula=I( V2 + V4) ~ I(V5*(V1 + V2) ) + I( V5*(V3
+ V4) ) + I( (1-V5)*(V1 + V2)) + I( (1-V5)*(V3+V4)) - 1, family =
                                                        binomial, data=tonsilData )
noTonsilModel <-
```

```
    glm( formula = I( V2 + V4) ~ I(V1 + V2) + I(V3 + V4) - 1, family=binomial, data=tonsilData )

## Under tonsil Model:
## P( Y = 1 | X = ( 0, 0, 1, 0 ) ) = 1 - \pi_{00}^N
## P( Y = 1 | X = ( 0, 0, 0, 1 ) ) = \pi_{11}^N
## P( Y = 1 | X = ( 1, 0, 0, 0 ) ) = 1 - \pi_{00}^W
## P( Y = 1 | X = ( 0, 1, 0, 0 ) ) = \pi_{11}^W
pi00.N <- 1.0 -  1.0 /
    ( 1.0 + exp( -sum( tonsilModel$coefficients* c(0,0,1,0)) ) )

pi11.N <- 1.0 /
    ( 1.0 + exp( -sum( tonsilModel$coefficients* c(0,0,0,1)) ) )

pi00.W <- 1.0 -  1.0 /
    ( 1.0 + exp( -sum( tonsilModel$coefficients*c(1,0,0,0)) ) )

pi11.W <- 1.0 /
    ( 1.0 + exp( -sum( tonsilModel$coefficients*c(0,1,0,0)) ) )

print( "Under the alternate:" )
print( sprintf("pi00.N = %f, pi11.N = %f, pi00.W = %f, pi11.N = %f", pi00.N, pi11.N, pi00.W, pi11.W ) )

## P( Y = 1 | X = ( 1, 0 ) ) = 1 - \pi_{00}
## P( Y = 1 | X = ( 0, 1 ) ) = \pi_{11}

pi11 <- 1.0 / ( 1.0 + exp( -sum( noTonsilModel$coefficients*c(0,1)) ) )
pi00 <- 1.0 -  1.0 /
    ( 1.0 + exp( -sum( noTonsilModel$coefficients*c(1,0)) ) )

print( "Under the Null:" )
print( sprintf( "pi11 = %f, pi00 = %f", pi00, pi11 ) )

chiStat <- 2*(logLik( tonsilModel )[1] - logLik( noTonsilModel )[1] )
pValue <- pchisq( chiStat, 2, lower.tail = FALSE )
print( sprintf( "Using a Likelihood ratio test: p-value of: %f", pValue ) )
```

```
[1] "Under the alternate:"
[1] "pi00.N = 0.062500, pi11.N = 0.666667, pi00.W = 0.375000, pi11.N = 0.363636"
[1] "Under the Null:"
[1] "pi11 = 0.250000, pi00 = 0.500000"
[1] "Using a Likelihood ratio test: p-value of: 0.003527"
```

We find that under the alternate we get very clean numbers for our probability estimates:

$$\begin{array}{ll} \pi_{00}^{\mathrm{N}} & \frac{15}{16} \\ \pi_{11}^{\mathrm{N}} & \frac{1}{3} \\ \pi_{00}^{\mathrm{W}} & \frac{5}{8} \\ \pi_{11}^{\mathrm{W}} & \frac{7}{11} \end{array}$$

Under the null hypothesis we get:

$$\begin{array}{ll} \pi_{00} & \frac{3}{4} \\ \pi_{11} & \frac{1}{2} \end{array}$$

1.10 j

The null hypothesis is given in part f. The probability that the null hypothesis is true is the p-value given as: 0.003527. This means that at the 5% confidence level, we have evidence to reject the null hypothesis. According to the data generated, there is a difference between the dynamics of throat infections based on tonsilectomies.