BDA - Assignment 2

Anonymous

Contents

Loaded packages

```
library(aaltobda)
library(stats)
library(tinytex)
```

Loaded data

```
data("algae")
```

```
NROW(algae) #this is n
```

```
## [1] 274
```

```
sum(algae) #this is y
```

```
## [1] 44
```

a)

Formulate model likelihood $p(y|\pi)$

We have: $n = 274$, $y = 44$. Thus:
$p(y|\pi) = Bin(y|n, \pi) = \binom{n}{y}\pi^y(1-\pi)^{n-y} = \binom{274}{44}\pi^{44}(1-\pi)^{230}$

Formulate the prior $p(\pi)$

$p(\pi) = Beta(\pi|\alpha, \beta) = Beta(\pi|2, 10)$ where $\alpha = 2$, $\beta = 10$

Formulate the resulting posterior $p(\pi|y)$

The beta prior distribution is a conjugate prior, thus the posterior distribution will have the same form as the prior distribution. As per BDA3 page 35, we thus have:

$p(\pi|y) = Beta(\pi|\alpha + y, \beta + n - y)$

We have: $\alpha = 2$, $\beta = 10$, $n = 274$, $y = 44$. Thus:

$p(\pi|y) = Beta(\pi|\alpha + y, \beta + n - y) = Beta(\pi|2 + 44, 10 + 274 - 44) = Beta(\pi|46, 240)$

b)

What can you say about the value of the unknown π according to the observations and your prior knowledge? Summarize your results with a point estimate (i.e. $E(\pi|y)$) and a 90% posterior interval).

The beta prior distribution is a conjugate prior, thus the posterior distribution will have the same form as the prior distribution. As per BDA3 page 35, we thus have:

With $\alpha = 2$, $\beta = 10$, $n = 274$, $y = 44$

$E[\pi|y] = \frac{\alpha+y}{\alpha+\beta+n} = \frac{2+44}{2+10+274} = \frac{46}{286} = 0.1608392$

```
46/286
```

```
## [1] 0.1608392
```

```
beta_point_est <- function(prior_alpha, prior_beta, data){
  n=NROW(algae)
  y=sum(algae)
  (prior_alpha+y)/(prior_alpha+prior_beta+n)}
beta_point_est_210=beta_point_est(prior_alpha = 2, prior_beta = 10, data)
beta_point_est_210
```

```
## [1] 0.1608392
```

This mean, which is approximately 0.1608392, is the posterior expectation of the parameter π.

To estimate the 90% posterior interval, we need to solve $P(\pi < c|y) = 0.05$ and $P(\pi > d|y) = 0.05$ for c and d.

```
prior_alpha = 2
prior_beta = 10
n=NROW(algae)
y=sum(algae)
posterior_alpha=prior_alpha+y
posterior_beta=prior_beta+n-y
beta_interval210=qbeta(c(0.05, 0.95), shape1=posterior_alpha, shape2=posterior_beta)
beta_interval210
```

```
## [1] 0.1265607 0.1978177
```

The 90% posterial interval is [0.1265607, 0.1978177]. In other words, according to the observations and prior knowledge, we can say that there is a 90% probability that the probability of a monitoring site having detectable blue-green algae levels (unknown π) is between 0.1265607 and 0.1978177.

c)

What is the probability that the proportion of monitoring sites with detectable algae levels π is smaller than $\pi_0 = 0.2$ that is known from historical records?

The cumulative density function is used to calculate the probability that the proportion of monitoring sites with detectable algae levels π is smaller than $\pi_0 = 0.2$ as follows:

```
pbeta(0.2,posterior_alpha,posterior_beta)
```

```
## [1] 0.9586136
```

The probability that the proportion of monitoring sites with detectable algae levels π is smaller than $\pi_0 = 0.2$ that is known from historical records $Pr(\pi < 0.2|y)$ is approximately 0.9586136.

d)

What assumptions are required in order to use this kind of a model with this type of data?

We have π be the probability of a monitoring site having detectable blue-green algae levels and y the observations in the dataset, algae. To use this kind of a model with this type of data, the required assumption is that the algae status at each site is an independent event and have the same distribution. Also, the observations y use a binomial model.

e)

Make prior sensitivity analysis by testing a couple of different reasonable priors and plot the different posteriors. Summarize the results by one or two sentences.

Prior Sensitivity Analysis

The tesing priors are: Beta(2,10) - original, Beta(5,10), Beta(10,10) and Beta(30,30).

- Estimates of Beta(2, 10) prior for π

```
#posterior mean
beta_point_est_210
```

```
## [1] 0.1608392
```

```r
#90% posterior interval
beta_interval210
```

```
## [1] 0.1265607 0.1978177
```

- Estimates of Beta(5, 10) prior for π

```r
# posterior mean
beta_point_est_510=beta_point_est(prior_alpha = 5, prior_beta = 10, data)
beta_point_est_510
```

```
## [1] 0.1695502
```

```r
#90% posterior interval
prior_alpha = 5
prior_beta = 10
n=NROW(algae)
y=sum(algae)
posterior_alpha=prior_alpha+y
posterior_beta=prior_beta+n-y
beta_interval510=qbeta(c(0.05, 0.95), shape1=posterior_alpha, shape2=posterior_beta)
beta_interval510
```

```
## [1] 0.1346454 0.2070585
```

- Estimates of Beta(10, 10) prior for π

```r
#posterior mean
beta_point_est_1010=beta_point_est(prior_alpha = 10, prior_beta = 10, data)
beta_point_est_1010
```

```
## [1] 0.1836735
```

```r
#90% posterior interval
prior_alpha = 10
prior_beta = 10
n=NROW(algae)
y=sum(algae)
posterior_alpha=prior_alpha+y
posterior_beta=prior_beta+n-y
beta_interval1010=qbeta(c(0.05, 0.95), shape1=posterior_alpha, shape2=posterior_beta)
beta_interval1010
```

```
## [1] 0.1478468 0.2219502
```

- Estimates of Beta(30, 30) prior for π

```r
#posterior mean
beta_point_est_3030=beta_point_est(prior_alpha = 30, prior_beta = 30, data)
beta_point_est_3030
```

```
## [1] 0.2215569
```

```
#90% posterior interval
prior_alpha = 30
prior_beta = 30
n=NROW(algae)
y=sum(algae)
posterior_alpha=prior_alpha+y
posterior_beta=prior_beta+n-y
beta_interval3030=qbeta(c(0.05, 0.95), shape1=posterior_alpha, shape2=posterior_beta)
beta_interval3030
```

```
## [1] 0.1852003 0.2598118
```

- Estimates Summary

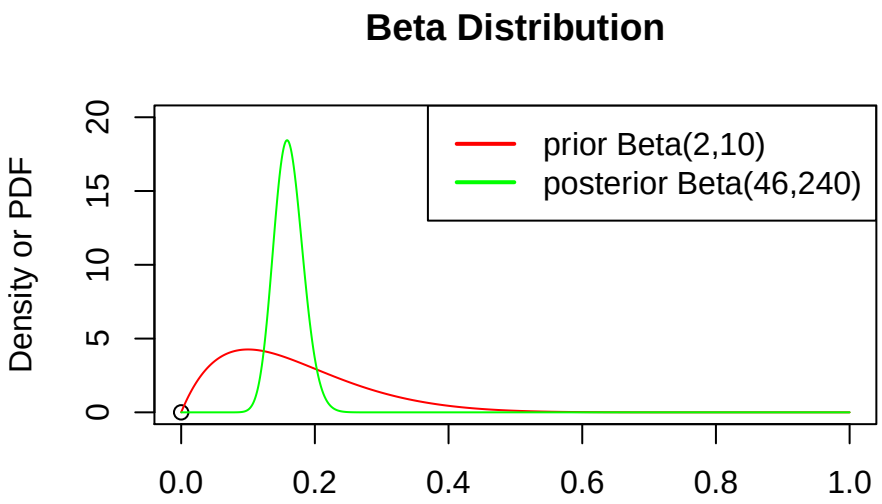
```
sensitivity_value <- matrix(cbind(round(beta_point_est_210,7),
                                  round(beta_point_est_510,7),
                                  round(beta_point_est_1010,7),
                                  round(beta_point_est_3030,7),
                                  '[0.1265607, 0.1978177]','[0.1346454, 0.2070585]',
                                  '[0.1478468, 0.2219502]', '[0.1852003, 0.2598118]'),
                            ncol=2)
colnames(sensitivity_value) <- c('Posterior Mean','90% Posterior Interval')
rownames(sensitivity_value) <- c('Beta(2,10)','Beta(5,10)','Beta(10,10)', 'Beta(30,30)')
sensitivity_value.table <- as.table(sensitivity_value)
sensitivity_value.table
```

```
##             Posterior Mean 90% Posterior Interval
## Beta(2,10)  0.1608392      [0.1265607, 0.1978177]
## Beta(5,10)  0.1695502      [0.1346454, 0.2070585]
## Beta(10,10) 0.1836735      [0.1478468, 0.2219502]
## Beta(30,30) 0.2215569      [0.1852003, 0.2598118]
```

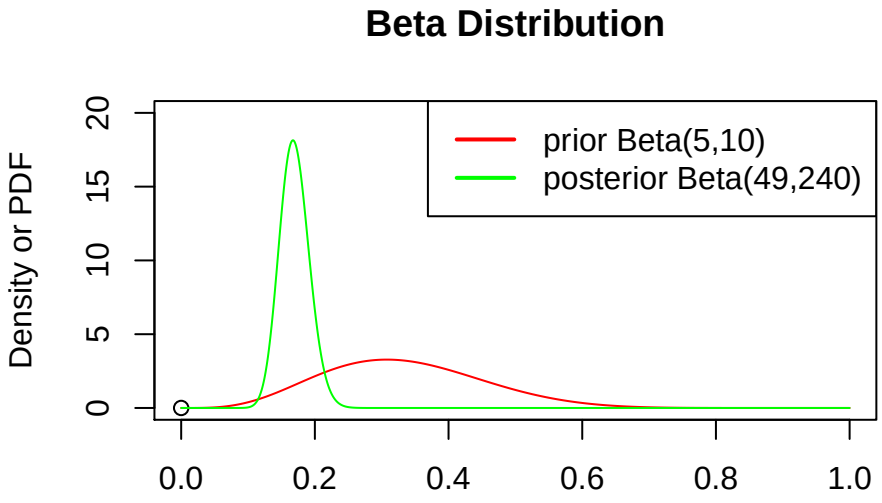
We can see that by varying the beta prior distribution significantly, we do not see extremely large changes in the posterior estimates. In other words, the poterior estimates are not extremely highly sensitive to the selection of priors. This may happen because of the high number of observation of the data set. Overall, we can say that posterior inferences based on a large sample are not particularly sensitive to the prior distribution.

- Plotting Summary

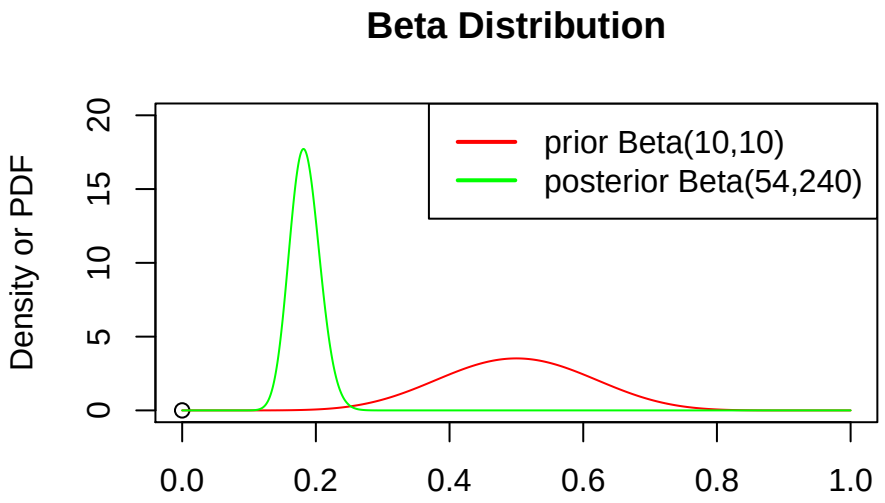
```
x<-seq(0,1,length.out=10000)
plot(0,0,main='Beta Distribution',xlim=c(0,1),ylim=c(0,20),ylab='Density or PDF',
     xlab='')
lines(x,dbeta(x,2,10),col='red')
lines(x,dbeta(x,46,240),col='green')
legend('topright',legend =c('prior Beta(2,10)', 'posterior Beta(46,240)'),
       col=c('red','green'),lwd=2)
```



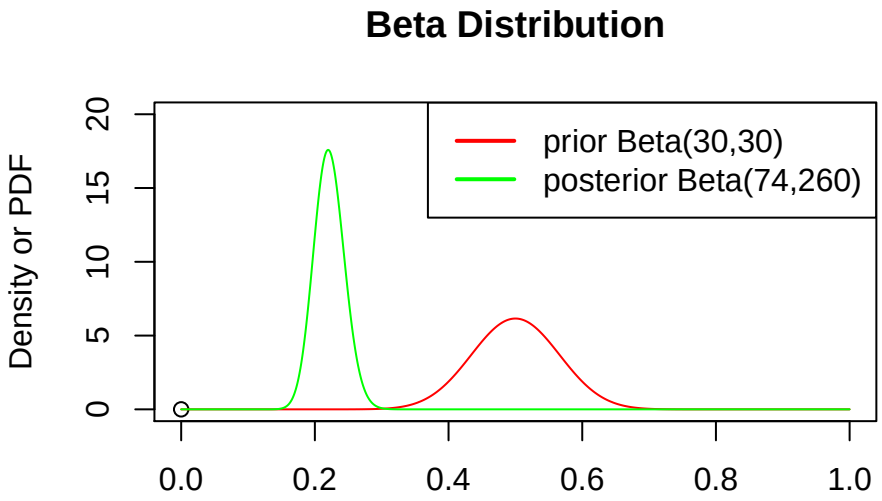

```
x<-seq(0,1,length.out=10000)
plot(0,0,main='Beta Distribution',xlim=c(0,1),ylim=c(0,20),ylab='Density or PDF',
     xlab='')
lines(x,dbeta(x,5,10),col='red')
lines(x,dbeta(x,49,240),col='green')
legend('topright',legend =c('prior Beta(5,10)', 'posterior Beta(49,240)'),
       col=c('red','green'),lwd=2)
```



```
x<-seq(0,1,length.out=10000)
plot(0,0,main='Beta Distribution',xlim=c(0,1),ylim=c(0,20),ylab='Density or PDF',
     xlab='')
lines(x,dbeta(x,10,10),col='red')
lines(x,dbeta(x,54,240),col='green')
legend('topright',legend =c('prior Beta(10,10)', 'posterior Beta(54,240)'),
       col=c('red','green'),lwd=2)
```




```
x<-seq(0,1,length.out=10000)
plot(0,0,main='Beta Distribution',xlim=c(0,1),ylim=c(0,20),ylab='Density or PDF',
     xlab='')
lines(x,dbeta(x,30,30),col='red')
lines(x,dbeta(x,74,260),col='green')
legend('topright',legend =c('prior Beta(30,30)', 'posterior Beta(74,260)'),
       col=c('red','green'),lwd=2)
```



From the graphs, posterior inferences based on a large sample are not particularly sensitive to the prior distribution.