Homework2b

yi Chen

9/14/2018

Homework 2b

Yi Chen

YC3356

yc3356@columbia.edu (mailto:yc3356@columbia.edu)

Exercise 2.22 of BDA

Parameter θ is the average improvement in success probability. Thus, here I make the model about this problem as a logistic regression problem. Note:

$$log(\frac{y}{1-y}) = \beta_0 + \theta x + error$$

where $x = 0, in\ the\ first\ month$ and $x = 1, in\ the\ second\ month\ after\ training,$

y is the success probability.

(a) noninformative prior

The improvement must within 0 and 1 because it is a probability. Thus, the noninformative prior could be in the following three conditions:

1. on logit scale: we can let $p(logit(\theta)) \sim constant$ which corresponds to the improper Beta(0,0). But if success probability is 0 or 1, the posterior woule be improper.
2. on probabilty scale: we can let $p(\theta) \sim constant$ which corresponds to the Beta(1,1)
3. we can also use the Jeffery's invariance principle to find the noninformative prior but these the likelihood here is the logit function of parameters which is relatively complicated, I do not try in this way.

Usually, as what is described in the textbook, the difference is small.

(b) subjective prior

Assume I know that, probably based on my personal experience or some research I read, I know that in this condition, there will have approximately 10% improvement in success probability, and I am very sure that this believe is very reasonable.

Thus, I could give a relatively strong subjective prior (smaller prior standard deviation), say $\theta \sim Beta(\alpha = 2, \beta = 18)$. Thus, the mean of prior is 0.1 and standard deviation is about 0.065

(c) weakly informative prior

Again, I know that, probably based on my personal experience or some research I read, I know that in this condition, there will have approximately 10% improvement in success probability. But I am not sure whether this believe is reasonable.

Thus, I could give a relatively strong subjective prior (smaller prior standard deviation), say $\theta \sim Beta(\alpha = 0.002, \beta = 0.018)$. Thus, the mean of prior is 0.1 and standard deviation is about 0.297 (much bigger)

Exercise 3.11 of

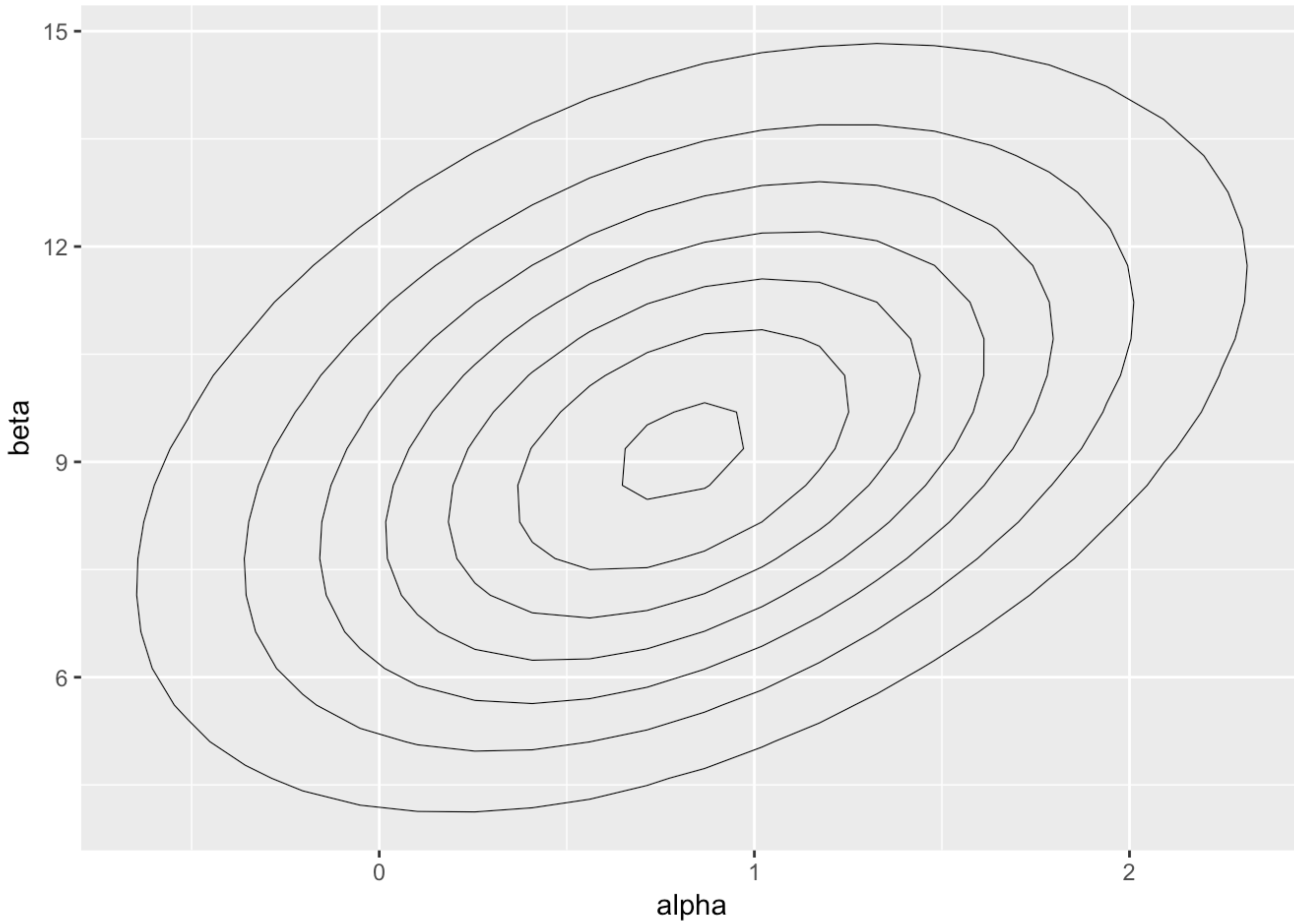
(a) repeat the computaions and plots in 3.7

step 1: set the data and function for calculate the unnormalizaed posterior probability for a given alpha and beta

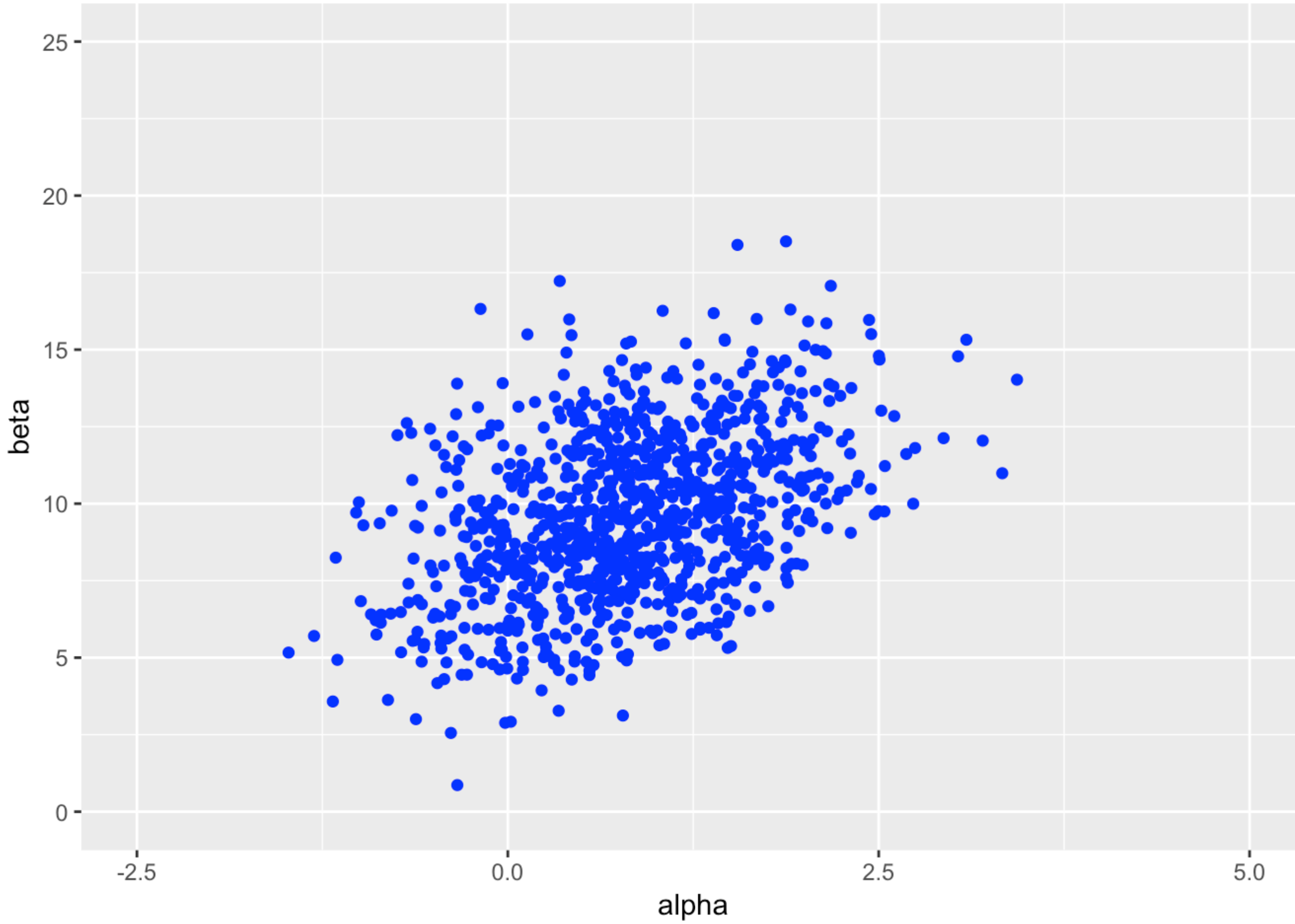
```
library(ggplot2)
library(mvtnorm)
library(gridExtra)
library(boot)
library(tidyr)


unnormalized_posterior <- function(alpha,beta){

  # prior function
    prior <- function(a=a,b=b){
        values <- c(a,b)
        # prior of alpha and beta
        mean_vector <- c(0,10)
        sigma_matric <- matrix(c(2,0.5,0.5,10),2,2)
        prob <- dmvnorm(values,mean = mean_vector,sigma=sigma_matric)
        return(prob)}

    # likelihood function
    likelihood <- function(a,b){
        ## bioassay data
        data <- data.frame(
          x = c(-0.86,-0.3,-0.05,0.73),
          n = c(5,5,5,5),
          y = c(0,1,3,5))
    # likelihood function based on the formula
        probs <- (inv.logit(a+b*data['x'][[1]])^data['y'][[1]])*((1-inv.logit(a+b*data['x'][[1]]))^(data['n'][[1
]]-data['y'][[1]]))
        result <- prod(probs)
        return(result)}

    #  calculate the posterior probability (vectorize method to improve the speed)
    posterior  <- mapply(prior,alpha,beta) * mapply(likelihood,alpha,beta)

    return(posterior)
}
```

```
## sample from grid
A = seq(-2.5, 5, length.out = 50)
B = seq(0, 25, length.out = 50)
cA <- rep(A, each = length(B))
cB <- rep(B, length(A))
nsamp <- 1000

p <- unnormalized_posterior(cA,cB)
samp_indices <- sample(length(p), size = nsamp,replace = T, prob = p/sum(p))
samp_A <- cA[samp_indices[1:nsamp]]
samp_B <- cB[samp_indices[1:nsamp]]
# Add random jitter
samp_A <- samp_A + runif(nsamp, (A[1] - A[2])/2, (A[2] - A[1])/2)
samp_B <- samp_B + runif(nsamp, (B[1] - B[2])/2, (B[2] - B[1])/2)
```

```
# Create a plot of the posterior density
# limits for the plots
xl <- c(-2.5, 5)
yl <- c(0, 25)
ggplot(data = data.frame(cA ,cB, p), aes(cA, cB, z= p)) +
  geom_contour(aes(z = p), colour = 'black', size = 0.2) +
  labs(x = 'alpha', y = 'beta') +
  scale_fill_gradient(low = 'yellow', high = 'red', guide = F) +
  scale_alpha(range = c(0, 1), guide = F)
```

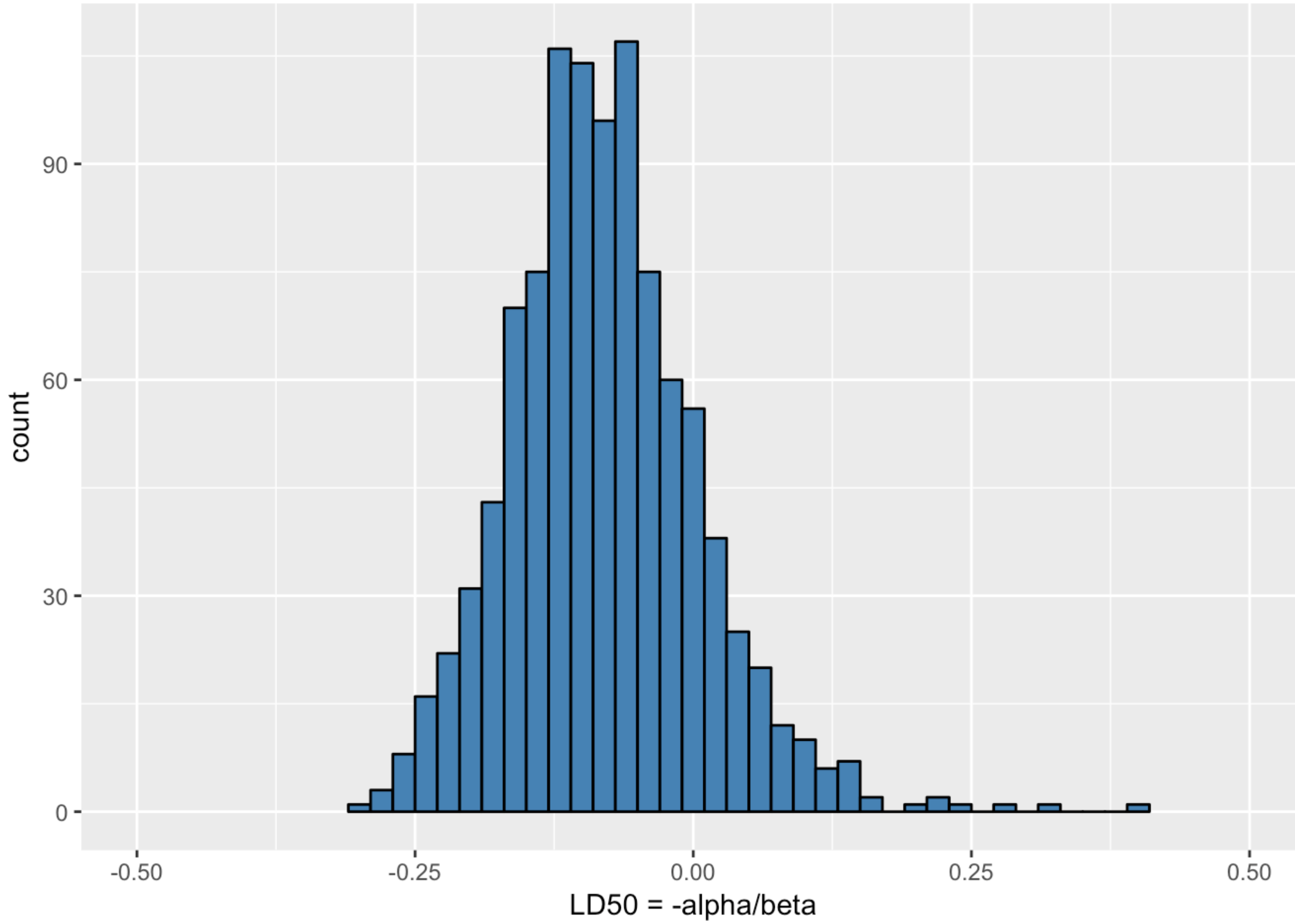



```
# Plot of the samples

ggplot(data = data.frame(samp_A, samp_B)) +
  geom_point(aes(samp_A, samp_B), color = 'blue') +
  coord_cartesian(xlim = xl, ylim = yl) +
  labs(x = 'alpha', y = 'beta')
```




```
# Sample LD50 conditional beta > 0
bpi <- samp_B > 0
samp_ld50 <- -samp_A[bpi]/samp_B[bpi]
```

```
# Plot of the histogram of LD50
ggplot() +
  geom_histogram(aes(samp_ld50), binwidth = 0.02,fill = 'steelblue', color = 'black') +
  coord_cartesian(xlim = c(-0.5, 0.5)) +
  labs(x = 'LD50 = -alpha/beta')
```



(b) check the contour plot

```
glm_data <- data.frame(
          x = rep(c(-0.86,-0.3,-0.05,0.73), each = 5),
          y = c(0,0,0,0,0,1,0,0,0,0,1,1,1,0,0,1,1,1,1,1))
logistic <- glm(data=glm_data,y~x,family = 'binomial')
summary(logistic)
```

```
##
## Call:
## glm(formula = y ~ x, family = "binomial", data = glm_data)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.37756  -0.64102  -0.07708   0.05473   1.83495
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.8466     1.0191   0.831    0.406
## x             7.7488     4.8727   1.590    0.112
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 27.526  on 19  degrees of freedom
## Residual deviance: 11.789  on 18  degrees of freedom
## AIC: 15.789
##
## Number of Fisher Scoring iterations: 7
```

As we can see from the result of logistic regression which gives us the esitmation based on likelihood only.

```
mean(samp_A)
```

```
## [1] 0.8162753
```

```
median(samp_A)
```

```
## [1] 0.8092617
```

```
mean(samp_B)
```

```
## [1] 9.627809
```

```
median(samp_B)
```

```
## [1] 9.608232
```

1. For α: prior mean is 0 and the logistic regression esimation is 0.8466. The posterior mean is 0.805 and the posterior median is 0.784. This indecate that it is a compromise between prior and likelihood.
2. For β: as the similar idea, the prior mean is 10 and logistic regression esitmation is 7.748. The posterior mean is 9.542 and the posterior median is 9.52. This indecate that it is a compromise between prior and likelihood.

(c) discuss the effect of the hyphothetical prior information on the conclusion in the applied context

The effect of prior information is different for two parameters.

For α the prior esimation shift to the logistic regression estimation more than prior mean. Given that the sample size is small, this means more weight is given to the likelihood and the informaion from prior is relativly small.

For β the the prior esimation shift to the prior mean more than logistic regression estimation. This means more weight is given to the data and the informaion from data is relativly small.