Homework 2 Sample Solutions

Courtesy of Chun-Chen Tu and Ran Bi

Oct 16, 2017

```r
# input the dataset
df = data.frame(year=1976:1985,
                fatalAccidents=c(24, 25, 31, 31, 22, 21, 26, 20, 16, 22),
                passengerDeaths=c(734, 516, 754, 877, 814, 362, 764, 809, 223, 1066),
                deathRate=c(0.19, 0.12, 0.15, 0.16, 0.14, 0.06, 0.13, 0.13, 0.03, 0.15))
df
```

```
##    year fatalAccidents passengerDeaths deathRate
## 1  1976             24             734      0.19
## 2  1977             25             516      0.12
## 3  1978             31             754      0.15
## 4  1979             31             877      0.16
## 5  1980             22             814      0.14
## 6  1981             21             362      0.06
## 7  1982             26             764      0.13
## 8  1983             20             809      0.13
## 9  1984             16             223      0.03
## 10 1985             22            1066      0.15
```

1

I would choose a prior distribution with mean about 20. We know that Gamma distribution is the conjugate prior for Poisson distribution. Thus, I would choose Gamma(α, β) where $\alpha = 40$, $\beta = 2$.

Let $y_1, ..., y_N$, denote the fatal accidents from 1976 to 1985 ($N = 10$) with Poisson(θ), the likelihood priot and the posterior distribution is:

$$p(y|\theta) \propto \prod_{n=1}^{N} \theta^{y_n} e^{-\theta} = \theta^{\sum_{n=1}^{N} y_n} e^{-n\theta} \tag{1}$$

$$p(\theta) \propto \theta^{\alpha-1} e^{-\beta\theta} \tag{2}$$

$$p(\theta|y) \propto \theta^{\alpha+\sum_{n=1}^{N} y_n - 1} e^{-\theta(\beta+N)} \tag{3}$$

And thus the posterior distribution follows Gamma($\alpha + \sum_{n=1}^{N} y_n$, $\beta + N$), which is Gamma(278, 12)

```r
alpha = 40
beta = 2
x = seq(0, 50, length.out=1000)

alpha_posterior = alpha + sum(df$fatalAccidents)
beta_posterior = beta + nrow(df)

# random sample from the posterior
nsample = 10000
theta = rgamma(nsample, alpha_posterior, beta_posterior)
```

```
obs = rpois(nsample, theta)
quantile(obs, c(0.025, 0.975))
```

```
##  2.5% 97.5%
##    14    33
```

The 95% predictive interval is [14, 33].

Alternatively, we may use the Jeffrey's prior:

$$J(\theta) \propto \sqrt{I(\theta)} = \sqrt{-E[\frac{d^2}{d\theta^2} log(\frac{e^{-\theta}\theta^y}{y!})|\theta]} = \sqrt{\theta^{-1}},$$

which is improper.

```
alpha = 1/2
beta = 0
x = seq(0, 50, length.out=1000)

alpha_posterior = alpha + sum(df$fatalAccidents)
beta_posterior = beta + nrow(df)

# random sample from the posterior
nsample = 10000
theta = rgamma(nsample, alpha_posterior, beta_posterior)
obs = rpois(nsample, theta)
quantile(obs, c(0.025, 0.975))
```

```
##  2.5% 97.5%
##    15    35
```

The 95% predictive interval in this case is [15, 35]. Result is relatively insensitive to the choice of prior. If your result differs significantly from this numerical value, it is likely that there is something wrong with your solution.

2

First we have to calculate the numbers of passenger miles in each year:

```
df = df %>% mutate(passengeMiles = passengerDeaths/deathRate*100*10e5)
df
```

```
##    year fatalAccidents passengerDeaths deathRate passengeMiles
## 1  1976             24             734      0.19  386315789474
## 2  1977             25             516      0.12  430000000000
## 3  1978             31             754      0.15  502666666667
## 4  1979             31             877      0.16  548125000000
## 5  1980             22             814      0.14  581428571429
## 6  1981             21             362      0.06  603333333333
## 7  1982             26             764      0.13  587692307692
## 8  1983             20             809      0.13  622307692308
## 9  1984             16             223      0.03  743333333333
## 10 1985             22            1066      0.15  710666666667
```

Let x_n be the number of passenger flown in year n. The posterior distribution is: Gamma($\alpha + \sum_{n=1}^{N} y_n$, $\beta + \sum_{n=1}^{N} x_n$), which is Gamma(278, 5.7159$\times 10^{12}$)

```r
alpha = 40
beta = 2
alpha_posterior = alpha + sum(df$fatalAccidents)
beta_posterior = beta + sum(df$passengeMiles)
nsample = 10000
theta = rgamma(nsample, alpha_posterior, beta_posterior)
obs = rpois(nsample, theta*8e11)
quantile(obs, c(0.025, 0.975))
```

```
##  2.5% 97.5%
##    27    53
```

The 95% predictive interval is [27, 53].

3

We choose the Jeffreys prior (Gamma distribution with $\alpha = 1/2$, $\beta = 0$)

```r
alpha = 1/2
beta = 0
x = seq(1, 1200, length.out=1000)

alpha_posterior = alpha + sum(df$passengerDeaths)
beta_posterior = beta + nrow(df)

# random sample from the posterior
nsample = 10000
theta = rgamma(nsample, alpha_posterior, beta_posterior)
obs = rpois(nsample, theta)
quantile(obs, c(0.025, 0.975))
```

```
##  2.5% 97.5%
##   639   746
```

The 95% predictive interval is [639, 746].

4

We choose the Jeffreys prior (Gamma distribution with $\alpha = 1/2$, $\beta = 0$)

```r
alpha = 1/2
beta = 0
alpha_posterior = alpha + sum(df$passengerDeaths)
beta_posterior = beta + sum(df$passengeMiles)
nsample = 10000
theta = rgamma(nsample, alpha_posterior, beta_posterior)
obs = rpois(nsample, theta*8e11)
quantile(obs, c(0.025, 0.975))
```

```
##  2.5% 97.5%
##   905  1032
```

The 95% predictive interval is [905, 1032].

5

The number of accidents data is more amenable to Poisson modeling, while the number of deaths is more appropriately modeled by a Compound Poisson Process.

Comparing a and b, it seems that the latter is more realistic in that the rate of accidents is proportional to passenger miles flown. Ignoring advances in safty technology over the years, it is natural to assume that the more passenger miles flown, the more accidents tend to happen.