BayesLinearRegress_cricket

2023-10-26

```r
library(MASS) #for mvrnorm
```

Bayesian Linear Regression. Cricket chirps example.

The basic idea of the linear regression is to understand the relationship between response y and explanatory variables $x = (x_1, ..., x_k)$, based on data from experimental units index by i. In general if we assume linearity between the mean of y and the explanatory variables, independence between experimental units, and constant variance around the mean, then we have:

$$y_i \overset{\text{iid}}{\sim} N(\beta_1 x_{i1} + ... + \beta_k x_{ik}, \sigma^2)$$

where $x_{i1} = 1$ if we want to include and intercept

In matrix notation:

$$y \sim N(X\beta, \sigma^2 I)$$

where $y = (y_1, ..., y_n)'$, $\beta = (\beta_1, ..., \beta_k)'$, and X is $n \times k$ with each row being $x_i = (x_{i1}, ..., x_{ik})$

Bayesian Regression:

Asumme the standard noninformative prior for joint distribution:

$$p(\beta, \sigma^2) \quad \alpha \quad \frac{1}{\sigma^2}$$

then the joint distribution posterior is

$$p(\beta, \sigma^2|y) = p(\beta|\sigma^2, y)p(\sigma^2|y)$$

where $p(\beta|\sigma^2, y)$ is the posterior distribution of β conditional to σ^2 and y, and $p(\sigma^2|y)$ the marginal posterior of σ^2.

$$\beta|\sigma^2, y \sim N(\hat{\beta}, \sigma^2 V_\beta)$$

$$\sigma^2|y \sim Inv - Gamma(\frac{n-k}{2}, \frac{(n-k)s^2}{2})$$

$$\beta|y \sim t_{n-k}(\hat{\beta}, s^2 V_\beta)$$

Hiperparameters:

$$\hat{\beta} = (X'X)^{-1}X'y$$

$$V_\beta = (X'X)^{-1}$$

$$s^2 = \frac{1}{n-k}(y - X\hat{\beta})'(y - X\hat{\beta})$$

The posterior is proper if $n > k$ and $rank(X) = k$ at least.

For numerical stability and efficiency, the QR decomposition should be used to calculate the posterior quantities.

Def. For a X, the QR decomposition is $X = QR$ for an orthogonal matrix Q and an upper triangular matrix R.

The quantity of interest are:

$$V_\beta = (X'X)^{-1} = ([QR]'QR)^{-1} = (R'Q'QR)^{-1} = (R'R)^{-1} = R^{-1}[R']^{-1}$$

$$\hat{\beta} = (X'X)^{-1}X'y = R^{-1}[R']^{-1}R'Q'y = R^{-1}Q'y =$$

$$R\hat{\beta} = Q'y$$

The last equation is useful because R is upper triangular and therefore the system of linear equations can be solved without requiring the inverse of R.

As an example, consider the relationship between the number of cricket chirps and temperature.

```r
chirps<-c(20,16,19.8,18.4,17.1,15.5,14.7,17.1,15.4,16.2,15,17.2,16,17,14.1)
temp<-c(88.6,71.6,93.3,84.3,80.6,75.2,69.7,82,69.4,83.3,78.6,82.6,80.6,83.5,76.3)
```

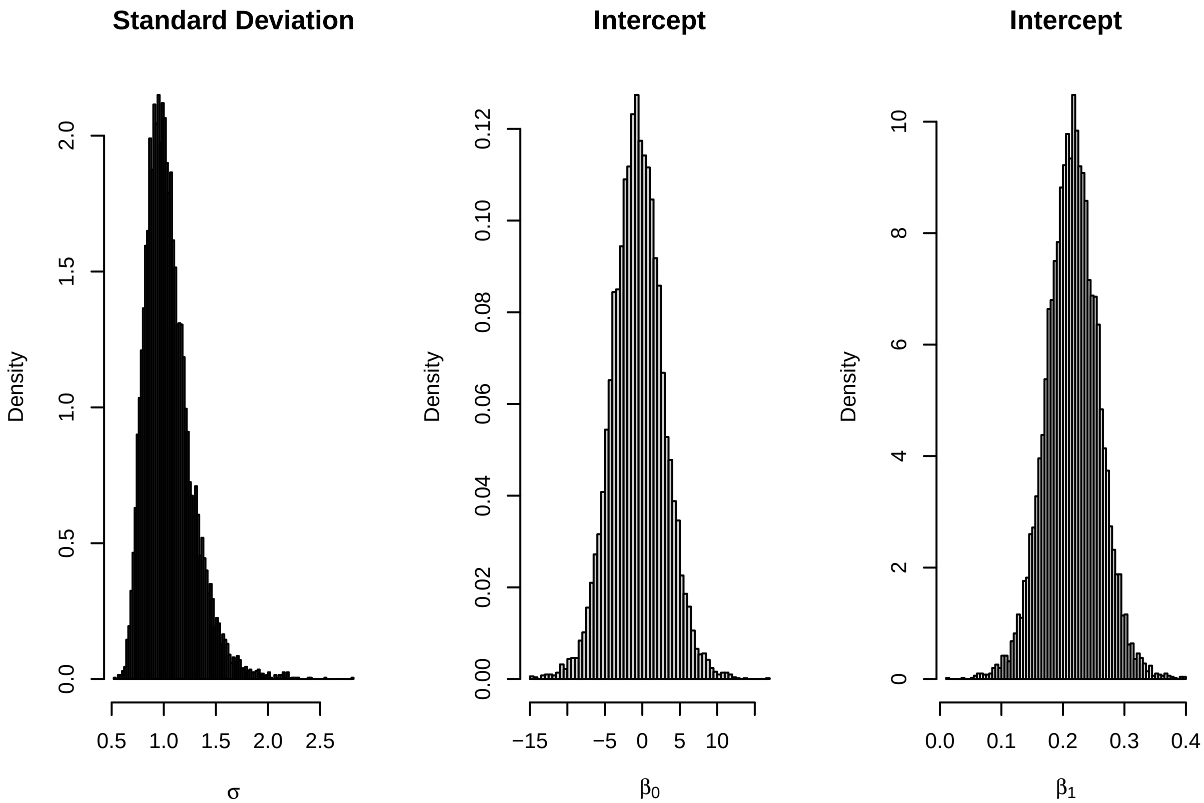
```r
X<-cbind(1,temp)
n<-nrow(X)
k<-ncol(X)
y<-matrix(chirps,n,1)

qr<-qr(X)
```

```r
#check for posterior propriety
stopifnot(n>k, qr$rank==k)
```

```r
#Calculate posterior hyperparameters
Rinv=solve(qr.R(qr))
vbeta<-Rinv%*%t(Rinv)
betahat<-qr.solve(qr,y)
df<-n-k
e=qr.resid(qr,y)
s2<-sum(e^2)/df
```

```r
#Simulte from the posterior
n.sims<-10000
sigma<-sqrt(1/rgamma(n.sims,df/2,df*s2/2))
beta<-matrix(betahat,n.sims,k,byrow=T)+sigma*mvrnorm(n.sims,rep(0,k), vbeta)
```

```r
par(mfrow=c(1,3))
hist(sigma,100,freq=F,main="Standard Deviation", xlab=expression(sigma))
hist(beta[,1],100,freq=F,main="Intercept", xlab=expression(beta[0]))
hist(beta[,2],100,freq=F,main="Intercept", xlab=expression(beta[1]))
```




```r
quantile(sigma, c(0.025,0.975))
```

```
##      2.5%     97.5%
## 0.7203271 1.5817162
```

```r
t(apply(beta,2,quantile,probs=c(0.025,0.975)))
```

```
##            2.5%     97.5%
##      -7.2777095 6.2485061
## temp  0.1299239 0.2996825
```

```r
confint(lm(y~0+X))
```

```
##            2.5 %    97.5 %
## X     -7.4081577 6.1777286
## Xtemp  0.1310169 0.3003406
```