STT 465

Variable Selection in Multiple Linear Regression

Variable selection in linear models

Regression equation

$$y_i = \sum_{j=1}^{p} x_{ij}\delta_j b_j + \varepsilon_i$$

$$\beta_j = \delta_j b_j$$

$$\delta_j \overset{iid}{\sim} Bernoulli(\pi)$$

$$b_j \overset{iid}{\sim} N\left(0, \sigma_b^2\right)$$

$\Rightarrow$ $\delta_j=0$ amounts to remove the j^{th} predictor from the model.

$\Rightarrow$ Next, we will discuss a Gibbs sampler for this model.

Variable selection in linear models

Likelihood

$$y_i = \sum_{j=1}^{p} x_{ij}\delta_j b_j + \varepsilon_i \qquad \varepsilon_i \overset{iid}{\sim} N\left(0,\sigma_\varepsilon^2\right)$$

$$p\left(y \middle| \delta, b, \sigma_\varepsilon^2\right) = \prod_{i=1}^{n} \frac{e^{-\frac{\left(y_i - \sum_j x_{ij}\delta_j b_j\right)^2}{2\sigma_\varepsilon^2}}}{\sqrt{2\pi\sigma_\varepsilon^2}}$$

Prior

$$p\left(\delta, b \mid \sigma_b^2, \pi\right) = \left[\prod_{j=1}^{p} \frac{e^{-\frac{b_j^2}{2\sigma_b^2}}}{\sqrt{2\pi\sigma_b^2}}\right] \times \left[\prod_{j=1}^{p} \pi^{\delta_j}\left(1-\pi\right)^{1-\delta_j}\right] \times p\left(\pi, \sigma_b^2, \sigma_\varepsilon^2\right)$$

Variable selection in linear models

Joint Posterior

$$p\left(\delta, b, \sigma_\varepsilon^2, \sigma_b^2, \pi \middle| y\right) \propto \left[\prod_{i=1}^{n} \frac{e^{-\frac{\left(y_i - \sum_j x_{ij}\delta_j b_j\right)^2}{2\sigma_\varepsilon^2}}}{\sqrt{2\pi\sigma_\varepsilon^2}}\right] \times \left[\prod_{j=1}^{p} \frac{e^{-\frac{b_j^2}{2\sigma_b^2}}}{\sqrt{2\pi\sigma_b^2}}\right]$$

$$\times \left[\prod_{j=1}^{p} \pi^{\delta_j} \left(1 - \pi\right)^{1-\delta_j}\right] \times p\left(\pi, \sigma_b^2, \sigma_\varepsilon^2\right)$$

Gibbs Sampler

Indicator Variables

$$p\left(\delta_k \middle| ELSE\right) \propto \left[\prod_{i=1}^{n} \frac{e^{-\frac{\left(y_i - \sum_{j \neq k} x_{ij}\delta_j b_j - x_{ik}\delta_k b_k\right)^2}{2\sigma_\varepsilon^2}}}{\sqrt{2\pi\sigma_\varepsilon^2}}\right] \pi^{\delta_k} \left(1-\pi\right)^{1-\delta_k}$$

$$p\left(\delta_k = 1 \middle| ELSE\right) \propto \left[\prod_{i=1}^{n} \frac{e^{-\frac{\left(y_i - \sum_{j \neq k} x_{ij}\delta_j b_j - x_{ik} b_k\right)^2}{2\sigma_\varepsilon^2}}}{\sqrt{2\pi\sigma_\varepsilon^2}}\right] \pi$$

$$p\left(\delta_k = 0 \middle| ELSE\right) \propto \left[\prod_{i=1}^{n} \frac{e^{-\frac{\left(y_i - \sum_{j \neq k} x_{ij}\delta_j b_j\right)^2}{2\sigma_\varepsilon^2}}}{\sqrt{2\pi\sigma_\varepsilon^2}}\right] \left(1-\pi\right)$$

Gibbs Sampler

Indicator Variables

$$p\left(\delta_k = 1 \middle| ELSE\right) = \frac{\left[\prod_{i=1}^{n} \frac{e^{-\frac{\left(y_i - \sum_{j\neq k} x_{ij}\delta_j b_j - x_{ik} b_k\right)^2}{2\sigma_\varepsilon^2}}}{\sqrt{2\pi\sigma_\varepsilon^2}}\right]\pi}{\left[\prod_{i=1}^{n} \frac{e^{-\frac{\left(y_i - \sum_{j\neq k} x_{ij}\delta_j b_j - x_{ik} b_k\right)^2}{2\sigma_\varepsilon^2}}}{\sqrt{2\pi\sigma_\varepsilon^2}}\right]\pi + \left[\prod_{i=1}^{n} \frac{e^{-\frac{\left(y_i - \sum_{j\neq k} x_{ij}\delta_j b_j\right)^2}{2\sigma_\varepsilon^2}}}{\sqrt{2\pi\sigma_\varepsilon^2}}\right](1-\pi)}$$

Gibbs Sampler

Effects

$$p\left(b_k \middle| ELSE\right) \propto \left[\prod_{i=1}^{n} \frac{e^{-\frac{\left(y_i - \sum_{j\neq k} x_{ij}\delta_j b_j - x_{ik}\delta_k b_k\right)^2}{2\sigma_\varepsilon^2}}}{\sqrt{2\pi\sigma_\varepsilon^2}}\right] \times \left[\frac{e^{-\frac{b_k^2}{2\sigma_b^2}}}{\sqrt{2\pi\sigma_b^2}}\right]$$

Case 1 ($\delta_k = 0$)

$$p\left(b_k \middle| ELSE\right) \propto \left[\frac{e^{-\frac{b_k^2}{2\sigma_b^2}}}{\sqrt{2\pi\sigma_b^2}}\right] = N\left(b_k \middle| 0, \sigma_b^2\right)$$

Case 2 ($\delta_k = 1$)

$$p\left(b_k \middle| ELSE\right) \propto \left[\prod_{i=1}^{n} \frac{e^{-\frac{\left(y_i - \sum_{j\neq k} x_{ij}\delta_j b_j - x_{ik} b_k\right)^2}{2\sigma_\varepsilon^2}}}{\sqrt{2\pi\sigma_\varepsilon^2}}\right] \times \left[\frac{e^{-\frac{b_k^2}{2\sigma_b^2}}}{\sqrt{2\pi\sigma_b^2}}\right]$$

(same as in the linear model without indicator variables)

Gibbs Sampler

Other variables

- Variances have the same fully conditionals as those of the standard multiple linear regression model.

- If we assign either a beta prior to π, the fully conditional can be shown to be beta.

Example

- **In this example we simulate data using 1000 predictors, out of which only 10 have effects and fit the variable selection model using BGLR.**

```
library(BGLR) ; data(mice)
X=mice.X[,1:1000]
QTL=seq(from=50,to=950,length=10)
b=rep(0,ncol(X)); b[QTL]<-1
signal<-X%*%b
error=rnorm(n=nrow(X),sd=sd(signal))
y=error+signal


fm=BGLR(y=y,
       ETA=list(list( X=X,model='BayesC',saveEffects=T))
       ,nIter=12000,burnIn=200)

plot(fm$ETA[[1]]$b,cex=.5,type='o',col=2,
       ylab='Estimated Effects');abline(v=QTL,col=4)

plot(fm$ETA[[1]]$d,cex=.5,type='o',col=2, ylab='p(dj=1)');
      abline(v=QTL,col=4)
```