Modelagem Bayesiana e Aplicações

Márcia D'Elia Branco

Universidade de São Paulo
Instituto de Matemática e Estatística
http:www.ime.usp.br/m̃branco

Modelos de Regressão

O modelo linear normal homocedástico

$$y = X\beta + \epsilon$$

$y = (y_1, \dots, y_n)^t$ é um vetor aleatório $n \times 1$ de variáveis respostas, X uma matrix de valores fixados $n \times p$ (variáveis explicativas) e $\beta = (\beta_1, \dots, \beta_p)^t$ é o vetor de parâmetros de dimensão $p \times 1$.

Além disso,

$$\epsilon \mid \sigma^2 \sim N_n(0, \frac{1}{\gamma} I_n)$$

condicionamente independentes.

O vetor de parâmetros a estimar é $\theta = (\beta, \gamma)$ em que γ é a precisão.

Regressão linear normal: Estimação

A função de verossimilhança:

$$L(\theta; y) \propto (\gamma)^{n/2} \exp\left\{-\frac{\gamma}{2}(y - X\beta)^t(y - X\beta)\right\}$$

Considere os usuais estimadores clássicos:

$$\hat{\beta} = (X^tX)^{-1}X^ty \quad \text{e} \quad S^2 = (y - X\hat{\beta})^t(y - X\hat{\beta})$$

Podemos verificar que

$$(y - X\beta)^t(y - X\beta) = (y - X\hat{\beta})^t(y - X\hat{\beta}) + (\beta - \hat{\beta})^tX^tX(\beta - \hat{\beta}).$$

Resulta na seguinte função de verossimilhança

$$L(\theta; y) \propto (\gamma)^{n/2} \exp\left\{-\frac{\gamma}{2}[S^2 + (\beta - \hat{\beta})^t X^t X(\beta - \hat{\beta})]\right\}$$

Usando uma a priori conjugada para $h(\beta, \gamma)$, isto é,

$$\gamma \sim Gama\left(\frac{n_0}{2}, \frac{n_0 T_0}{2}\right) \quad \text{e} \quad \beta \mid \gamma \sim N_p\left(c_0, \frac{C_0}{\gamma}\right)$$

Obtemos que a distribuição a posteriori é proporcional a

$$(\gamma)^{\frac{n+p+n_0}{2}-1} \exp\left\{-\frac{\gamma}{2}\left[n_0 T_0 + S^2 + (\beta - c_0)^t C_0^{-1}(\beta - c_0) + R(\beta)\right]\right\}$$

com $R(\beta) = (\beta - \hat{\beta})^t X^t X(\beta - \hat{\beta})$

Assim $h(\beta, \gamma \mid y)$ é proporcional a

$$(\gamma)^{\frac{p+n_1}{2}-1} \exp\left\{-\frac{\gamma}{2}\left[(\beta - c_1)^t C_1^{-1}(\beta - c_1) + n_1 T_1\right]\right\}.$$

Resultando em

$$\gamma \mid y \sim Gama\left(\frac{n_1}{2}, \frac{n_1 T_1}{2}\right) \quad \text{e} \quad \beta \mid \gamma, y \sim N\left(c_1, \frac{C_1}{\gamma}\right)$$

em que $n_1 = n_0 + n$ e

$$C_1^{-1} = C_0^{-1} + X^t X,$$

$$c_1 = C_1(C_0^{-1} c_0 + X^t y),$$

$$n_1 T_1 = n_0 T_0 + (y - Xc_1)^t y + (c_0 - c_1)^t C_0^{-1} c_0.$$

Regressão linear normal: Estimação

- A distribuição a posteriori marginal de β também pode ser obtida analiticamente, resultando em uma t-multivariada (ver Paulino et al., 2018).

$$\beta \mid y \sim T_p(c_1, T_1 C_1, n_1)$$

- Se considerarmos prioris independentes $\gamma \sim Gama(a_0, b_0)$ e $\beta \sim N_p(c_0, d_0)$ não obtemos o mesmo resultado de conjugação. No entanto, é possível obter as condicionais completas e aplicar a algoritmo de Gibbs (ver pag. 99 de Congdon).
- $\beta \mid y, \gamma$ será normal e $\gamma \mid y, \beta$ uma Gama.

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i.$$

Temos :

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \dots & \dots \\ 1 & x_n \end{bmatrix} \quad , \quad X^t X = \begin{bmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{bmatrix}$$

e

$$X^t y = \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} y_i x_i \end{bmatrix}.$$

Caso especial $p = 2$

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

E os estimadores de máxima verossimilhança são dados por:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \quad , \quad \hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) y_i}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

Considerando a priori de Jeffreys $h(\beta, \sigma^2) \propto 1/\sigma^2$ temos que

$$\beta \mid \gamma, y \sim N_2\left((\hat{\beta}_0, \hat{\beta}_1)^t, (\gamma X^t X)^{-1}\right) \text{ e } \gamma \mid y \sim Gama\left(\frac{n-2}{2}, \frac{S^2}{2}\right)$$

com

$$S^2 = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = (n-2)s_\epsilon^2$$

As estimativas pontuais para β são os emv:

$$E[\beta \mid y] = E[E[\beta \mid \gamma, y]] = \hat{\beta} \quad , n > 2.$$

A estimativa pontual para γ é

$$E[\gamma \mid y] = \frac{1}{s_\epsilon^2}.$$

- Os IC são facilmente obtidos, usando os quantis das distribuições a posteriori.
- Se centrarmos as covariáveis obtemos independência a posteriori para os β_j condicional a γ.
- Note que, $p = 2$, fazendo $x_i^* = x_i - \bar{x}$, temos que a matrix $X^t X$ fica diagonal. Portanto, β_0 independente de β_1 dado γ .
- A utilização de variáveis centradas também facilita na inversão da matriz $X^t X$.

Seja $Y_0 = X_0\beta + \epsilon$ com $\epsilon \sim N_q\left(0, \frac{1}{\gamma} I_q\right)$ um novo vetor.
X_0 é uma matriz $q \times p$ conhecida.

Sob a priori não informativa de Jeffreys, temos que a distribuição preditiva condicional é dada por

$$Y_0 \mid \gamma, y \sim N_q\left(X_0\hat{\beta}, \frac{1}{\gamma} C\right)$$

onde $C = I_q + X_0(X^tX)^{-1}X_0^t$.

Lembrando que $\gamma \mid y \sim Gama\left(\frac{n-p}{2}, \frac{S^2}{2}\right)$, resulta

$$Y_0 \mid y \sim T_q\left(X_0\hat{\beta}, s_\epsilon^2 C; n-p\right).$$

Adequação do modelo

- Para verificar a adequação do modelo podemos usar a preditiva a posteriori

$$f(y_{rep} \mid y) = \int f(y_{rep} \mid \theta, y) h(\theta \mid y) d\theta$$

- Uma possibilidade é considerar uma medida global de discrepância $V(y_{rep}, y_{obs})$ e obter o valor-P bayesiano.
- Alternativamente, podem ser avaliadas cada observação individualmente. Verificando se $I(y_{rep,i} > y_{obs,i})$ para cada $i = 1, \ldots, n$.
- Em ambos os casos, temos que simular réplicas da preditiva.

Para o modelo normal com priori de Jeffreys, temos:

$$\gamma \mid y \sim Gama\left(\frac{n-p}{2}, \frac{S^2}{2}\right)$$

$$\beta \mid \gamma, y \sim N_p\left(\hat{\beta}, [\gamma X^t X]^{-1}\right)$$

Considerando $y_{rep,i} \mid \theta \sim N(x_i^t \beta, 1/\gamma)$, temos que

$$y_{rep,i} \mid \gamma, y \sim N(x_i^t \hat{\beta}, \Delta/\gamma)$$

com

$$\Delta = 1 - x_i^t (X^t X)^{-1} x_i.$$

Neste caso, para simular uma réplicas de y_i basta considerar $\gamma^{(j)}$ simulado da *Gama* $\left(\frac{n-p}{2}, \frac{S^2}{2}\right)$ e simular $y_i^{(j)}$ da $N(x_i^t\hat{\beta}, \Delta/\gamma^{(j)})$.

No caso mais geral, proceder como antes, simulando amostra das posteriori de $\theta = (\beta, \gamma)$.

Vamos considerar

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}.$$

- O problema de seleção de variáveis é um problema similar ao de seleção de modelos. Portanto, as medidas usuais de comparação de modelo como, por exemplo, DIC, WAIC e CPO podem ser usadas.
- No entanto, temos 2^p possíveis modelos que pode ser um problema para valores altos de p. Usualmente é feita uma pré-seleção a priori que descarta alguns dos modelos não interessantes.

- Existem técnicas específicas de seleção de variáveis que adicionam ao modelo uma variável latente indicadora $\delta_j = 1$, se a j-ésima variável deve ser incluída no modelo e $\delta_j = 0$, caso contrário.
- Considerando uma probabilidade a priori $P(\delta_j = 1) = r_j$, obtem-se a probabilidade a posteriori $P(\delta_j = 1 \mid y)$.
- Essa probabilidade é obtida de forma aproximada via simulação da posteriori conjunta $h(\delta, \theta \mid y)$.
- O método SSVS (*stochastic search variable selection*) propõe uma priori hierárquica dada por

$$h(\beta_j \mid \delta_j) = (1 - \delta_j)N(0, c\tau_j^2) + \delta_j N(0, \tau_j^2) \qquad 0 < c < 1$$

$$\delta_j \sim Be(r_j)$$

- Uma proposta alternativa é o uso de prioris independentes $h(\beta_j, \delta_j) = h(\beta_j)h(\delta_j)$. Mantendo a priori Bernoulli para δ_j e

$$\beta_j \sim N(0, V_j).$$

- A inferência é baseada no controle da quantidade $\delta_j \beta_j$.
- Uma dica para ambas abordagens é considerar as variáveis padronizadas, isto é, subtrair a média e dividir pelo desvio-padrão amostrais.
- Extensões do método de mistura finita como priori para β_j (SSVS) são obtidos considerando-se misturas infinitas.

Considere

$$\beta_j \mid \delta_j \sim N(0, \sigma^2 \delta_j) \;\; \text{e} \;\; \delta_j \sim G(\delta_j).$$

- O método conhecido como Lasso Bayesiano é obtido fazendo $\delta_j \mid \lambda \sim Exp(\lambda^2/2)$.
- Neste caso, podemos mostrar que

$$h(\beta_j \mid \sigma^2, \lambda) = \frac{\lambda}{2\sigma} exp\{-\lambda|\beta_j|/\sigma\}.$$

- A distribuição acima é conhecida pelo nome de Laplace ou Exponencial Dupla.
- Se fizermos $\delta \mid \tau \sim Cauchy(0, \tau^2) I(0, \infty)$ (Cauchy truncada em valores positivos), temos o método conhecido como *Horseshoe* .