Statistique bayésienne
Feuille 4: Algorithme de Gibbs. Régression linéaire bayésienne.

Robin Ryder

Automne 2019

Nous allons effectuer de la régression linéaire dans le paradigme bayésien, pour expliquer le taux de mortalité dans des métropoles américaines dans les années 1960 [1].

Chargez les données `deathrate.csv`

Il y a 15 variables explicatives :

1. pluviométrie annuelle (en pouces)
2. température moyenne en janvier (en ^0F)
3. température moyenne en juillet (en ^0F)
4. proportion de la population âgée de plus de 65 ans
5. taille moyenne des foyers
6. nombre moyen d'années d'études
7. proportion de foyers avec une cuisine équipée
8. densité de population (par mile carré)
9. pourcentage de non blancs dans la population
10. pourcentage d'employés
11. pourcentage de familles pauvres (revenu inférieur à 3000 dollars/an)
12. pollution aux hydrocarbures
13. pollution aux oxydes de nitrogène
14. pollution au dioxyde de soufre
15. humidité moyenne

La 16ᵉ colonne donne le taux de mortalité.

Notre modèle linéaire est (avec $k = 15$)

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_i^1 + \beta_2 x_i^2 + \ldots + \beta_k x_i^k + \epsilon_i \\ \epsilon_i &\sim \mathcal{N}(0, \sigma^2) \end{aligned}$$

1. **Analyse fréquentiste** Donner les estimations fréquentistes de β et σ^2 à l'aide de la fonction `lm` ; on les note $\hat{\beta}$ et s^2.

1. Source des données : Gunst & Mason 1980, McDonals & Schwing 1973, Spaeth 1991.

2. **Inférence bayésienne à l'aide de la loi a priori g de Zellner** On considère la loi a priori suivante :

$$\begin{aligned} \beta|\sigma^2, X &\sim \mathcal{N}_{k+1}\left(0, g\sigma^2({}^tXX)^{-1}\right) \\ \pi(\sigma^2) &\propto \sigma^{-2} \end{aligned}$$

Cette loi est conjuguée ; la loi a posteriori associée est

$$\begin{aligned} \beta|\sigma^2, y, X &\sim \mathcal{N}_{k+1}\left(\frac{g}{g+1}\hat{\beta}, \frac{\sigma^2 g}{g+1}(X^TX)^{-1}\right) \\ \sigma^2|y, X &\sim \mathcal{IG}\left(\frac{n}{2}, \frac{s^2}{2} + \frac{1}{2(g+1)}\hat{\beta}^T X^T X\hat{\beta}\right) \end{aligned}$$

donc

$$\beta|y, X \sim \mathcal{T}_{k+1}\left(n, \frac{g}{g+1}\hat{\beta}, \frac{g(s^2 + \hat{\beta}^T X^T X\hat{\beta}/(g+1))}{n(g+1)}(X^TX)^{-1}\right).$$

   a) Pour $g = 0.1, 1, 10, 100, 1000$, donner $E[\sigma^2|y, X]$ et $E[\beta_0|y, X]$. Quel est l'impact de la loi a priori sur la loi a posteriori ?

   b) On souhaite tester l'hypothèse $H_0 : \beta_7 = \beta_8 = 0$.

   Pour une hypothèse H_0 à $k - q$ coefficients non nuls représentée par la matrice de covariables X_0, la vraisemblance marginale s'écrit

$$(g+1)^{-(k+1-q)/2}\pi^{-n/2}\Gamma(n/2) \times \left[y^Ty - \frac{g}{g+1}y^T X_0(X_0^T X_0)^{-1}X_0^T y\right]^{-n/2}$$

   Calculer le facteur de Bayes pour notre test et l'interpréter selon l'échelle de Jeffreys.

3. **Choix de modèle : calcul exact** Dans cette question, on se restreint aux 3 premières variables explicatives.

   On souhaite savoir quelles variables inclure dans notre modèle, et on fait l'hypothèse que $\beta_0 \neq 0$. Il y a donc $2^3 = 8$ modèles possibles. À chaque modèle on associe la variable $\gamma = (\gamma_1, \gamma_2, \gamma_3)$ où $\gamma_i = 1$ si x^i est incluse dans le modèle ($\beta_i \neq 0$) et $\gamma_i = 0$ sinon. Soit X_γ la sou-matrice de X dans laquelle on garde uniquement les colonnes i telles que $\gamma_i = 1$.

   À l'aide de la formule donnée dans la question précédente, calculer la vraisemblance marginale de chaque modèle. En déduire le modèle le plus probable a posteriori.

4. **Choix de modèle par échantillonnage de Gibbs** On considère maintenant toutes les variables explicatives. Il y a donc 2^k modèles parmi lesquels choisir. Échire un échantillonneur de Gibbs qui échantillonne selon la loi a posteriori de γ, et conclure sur le modèle le plus probable a posteriori. Le comparer à l'estimateur fréquentiste.

5. Pour une nouvelle valeur de votre choix des covariables, faire une prédiction à l'aide du modèle le plus probable, et à l'aide d'un mélange des modèles pondérés par leur probabilité a posteriori.