Devoir maison
Statistique bayésienne

Robin Ryder

Il est possible d'obtenir une très bonne note sans répondre à toutes les sections. Vous pouvez réutiliser le code écrit en cours en l'adaptant comme nécessaire. Pour certaines questions, plusieurs approches sont possibles. Lorsque vous implémentez un algorithme, expliquez brièvement comment vous avez vérifié qu'il a convergé.

Les enseignants des collèges et lycées français souhaitant obtenir une mutation professionnelle sont classés en fonction d'un nombre de points qui dépend de leur situation personnelle et de leur carrière. Le fichier `mutations2.csv` donne le nombre de points nécessaire pour obtenir une mutation dans les lycées de l'académie de Versailles en 2012, pour diverses disciplines enseignées ; c'est une mesure de l'attractivité de chaque établissement pour les enseignants. Par exemple, en mathématiques, il suffisait de 21 points pour pouvoir être nommé au lycée Georges Braque d'Argenteuil, mais il en fallait 464 pour être nommé au lycée Michelet de Vanves. Nous allons étudier ce nombre de points, dans un cadre bayésien.

Pour des couples (établissement, discipline), on dispose du nombre de points nécessaire (colonne `Barre`) pour obtenir une mutation, ainsi que de caractéristiques de l'établissement : nombre de candidats au baccalauréat par série, taux de réussite au baccalauréat par série, taux de réussite attendu (qui dépend notamment du tissu socioprofessionnel des parents d'élèves), taux d'accès des élèves de seconde et de première au baccalauréat. Par souci d'homogénéité des données, on considère uniquement les filières du lycée général, même si beaucoup des établissements concernés préparent aussi au baccalauréat technologique et parfois au baccalauréat professionnel.

1 Régression linéaire

On propose d'abord un modèle linéaire gaussien. On cherche à expliquer le nombre de points nécessaire à une mutation (colonne `Barre`) par les caractéristiques du lycée.

1. Effectuer une régression linéaire bayésienne et interpréter les coefficients obtenus.
2. Choisir les covariables significatives. Comparer au résultat obtenu par une analyse fréquentiste. *Afin de réduire le coût computationnel, il peut être intéressant d'effectuer une présélection des covariables considérées.*
3. On se concentre maintenant uniquement sur les mutations en mathématiques et en anglais. Répéter l'analyse pour chacune de ces deux catégories. Que penser de l'hypothèse que les covariables agissent de la même manière dans ces deux disciplines ?

2 Loi de Pareto

On ignore maintenant les covariables, et on s'intéresse uniquement à la loi du nombre de points nécessaire (colonne `Barre`). La loi gaussienne peut paraître peu pertinente pour ces données : on va

plutôt proposer une loi de Pareto. Pour $m > 0$ et $\alpha > 0$, on dit que $Z \sim Pareto(m, \alpha)$ si Z est à valeurs dans $[m, +\infty[$ de densité

$$f_Z(z; m, \alpha) = \alpha \frac{m^\alpha}{z^{\alpha+1}} \mathbb{I}_{\{z \geq m\}}.$$

On impose $m = 21$ au vu des données.

4. Chercher un package R permettant de générer des réalisations d'une loi de Pareto. Visualiser l'impact du paramètre α.
5. Choisir une loi a priori pour α.
6. Donner la loi a posteriori de α.
7. Par la méthode de votre choix, tirer un échantillon de la loi a posteriori de α. Donner un intervalle de crédibilité à 95%.
8. On se concentre uniquement sur les mutations en mathématiques et en anglais. Répéter l'analyse pour chacune de ces deux catégories. Que pensez-vous de l'hypothèse que $\alpha_{\text{maths}} = \alpha_{\text{anglais}}$?