TP7: régression logistique
Statistiques pour Sciences Humaines II

Baptiste Perez et Anna Kiriliouk

Année 2021–2022

NB : toutes les réponses ont été arrondies à deux chiffres après la virgule.

Réponses

1. Voir fichier `R`. En général, la personne semble un peu plus apprécier les chansons qui sont en mineur que ceux en majeur, et il/elle a une petite préférence pour les chansons plus dansantes.
2. Dans la suite, on écrit `dance` pour `danceability`. Le modèle linéaire estimé (pour le logarithme de la cote d'aimer une chanson) est

$$\widehat{\beta}_0 + \widehat{\beta}_{\texttt{mode}} \times \texttt{mode} + \widehat{\beta}_{\texttt{dance}} \times \texttt{dance} = -1.19 - 0.27 \times \texttt{mode} + 2.23 \times \texttt{dance}$$

   où `mode = 1` pour `majeur` et `mode = 0` pour `mineur`. La cote d'aimer une chanson pour chacun des deux modes est alors une fonction de `dance`,

$$\begin{aligned} \widehat{c}(\texttt{dance}) &= \exp\left(-1.46 + 2.23 \times \texttt{dance}\right), \qquad \texttt{(majeur)} \\ \widehat{c}(\texttt{dance}) &= \exp\left(-1.19 + 2.23 \times \texttt{dance}\right), \qquad \texttt{(mineur)} \end{aligned}$$

3. Pour un même niveau de `dance`, disons `dance = 0.35`, la cote d'aimer une chanson en majeur est $e^{\widehat{\beta}_{\texttt{mode}}} = e^{-0.27} \approx 0.76$ fois la cote d'aimer une chanson en mineur. Par exemple, on calcule

$$\widehat{c}(0.35) = \exp(-1.19 + 2.23 \times 0.35) \approx 0.66, \qquad \texttt{(mineur)}$$

   c'est-à-dire, la cote d'aimer une chanson en mineur est de $2:3$, et donc la cote d'aimer une chanson en majeur est de $0.76 \times \frac{2}{3} \approx 0.5$, c'est-à-dire, de $1:2$. En effet,

$$\widehat{c}(0.35) = \exp(-1.46 + 2.23 \times 0.35) \approx 0.5, \qquad \texttt{(majeur)}$$

4. Quand `dance` augmente de 0.35 à 0.45, ou de 0.45 à 0.55, que ce soit pour une chanson en mineur ou une chanson en majeur, la cote d'aimer la chanson est multipliée par $e^{0.1 \times \widehat{\beta}_{\texttt{dance}}} = e^{0.1 \times 2.23} \approx 1.25$, car

$$\frac{\widehat{c}(\texttt{dance + 0.1})}{\widehat{c}(\texttt{dance})} = \frac{\exp\left(\texttt{-1.19 - 0.27 mode + 2.23 (dance + 0.1)}\right)}{\exp\left(\texttt{-1.19 - 0.27 mode + 2.23 dance}\right)} = e^{0.1 \times 2.23}$$

5. On calcule $\widehat{\beta}_0/\widehat{\beta}_{\text{dance}} = -1.19/2.23 = -0.534$ et $(\widehat{\beta}_0 + \widehat{\beta}_{\text{mode}})/\widehat{\beta}_{\text{dance}} = -1.46/2.23 = -0.656$ pour écrire

$$\widehat{c}(\texttt{dance}) = \exp\left(2.23(-0.656 + \texttt{dance})\right), \qquad \texttt{(majeur)}$$
$$\widehat{c}(\texttt{dance}) = \exp\left(2.23(-0.534 + \texttt{dance})\right), \qquad \texttt{(mineur)}$$

La cote d'aimer une chanson en majeur est de 1 : 1 quand `dance = 0.656` et la cote d'aimer une chanson en mineur est de 1 : 1 quand `dance = 0.534`.

6. Voir fichier `R`. On voit que les fonctions de régression sont quasi linéaires. On peut retrouver la "forme en S" en regardant des valeurs de `danceability` entre -2 et 3 ; or, ces valeurs ne veulent rien dire ! On ne risque donc pas de prédire des probabilités en dehors de l'intervalle $[0, 1]$. La régression logistique n'ajoute donc pas grand chose par rapport à la régression linéaire dans cet exemple.
7. Voir fichier `R`.
8. On ne voit aucune corrélation élevée. Il n'y a donc pas de risque de multicolinéarité.
9. On enlève la variable `loudness` car la p-valeur associée est plus grande que 0.05. Le modèle linéaire estimé (pour le logarithme de la cote d'aimer une chanson) est

$$\begin{aligned}&\widehat{\beta}_0 + \widehat{\beta}_{\text{mode}} \times \texttt{mode} + \widehat{\beta}_{\text{dance}} \times \texttt{dance} + \widehat{\beta}_{\text{duration}} \times \texttt{duration} + \widehat{\beta}_{\text{tempo}} \times \texttt{tempo} + \widehat{\beta}_{\text{instru}} \times \texttt{instru}\\ &= -3.12 - 0.20 \times \texttt{mode} + 2.76 \times \texttt{dance} + 0.0029 \times \texttt{duration} + 0.0056 \times \texttt{tempo} + 1.20 \times \texttt{instru}\end{aligned}$$

10. Dans la question 7, on avait qu'un régresseur quantitatif. Maintenant on en a 4, et donc on devrait faire 4 graphiques différents : dans chaque graphique, on mettrait une des variables `duration, danceability, tempo, instrumentalness` sur l'abscisse. Pour calculer la courbe des probabilités estimées, dans chaque graphique, il faut choisir une valeur fixe pour les autres trois variables, par exemple, leur moyenne. Le fichier `R` contient un exemple d'un de ces graphiques (pour la probabilité d'aimer une chanson en fonction de sa durée). Vous ne devez pas être capables d'en faire vous mêmes.
11. La probabilité d'aimer la chanson "sign of the times", est, d'après les deux modèles (ci-dessous les coefficients estimés ont été arrondis)

$$\texttt{(m1)} \quad \widehat{p} = \frac{1}{1 + \exp\left(1.46 - 2.23 \times 0.516\right)} \approx 0.42$$
$$\texttt{(m2)} \quad \widehat{p} = \frac{1}{1 + \exp\left(3.33 - 0.0029 \times 340.7 - 2.76 \times 0.516 - 0.0056 \times 120 - 1.2 \times 0\right)} \approx 0.45$$

La probabilité d'aimer la chanson "Hotline Bling", est, d'après les deux modèles

$$\texttt{(m1)} \quad \widehat{p} = \frac{1}{1 + \exp\left(1.46 - 2.23 \times 0.896\right)} \approx 0.63$$
$$\texttt{(m2)} \quad \widehat{p} = \frac{1}{1 + \exp\left(3.33 - 0.0029 \times 267 - 2.76 \times 0.896 - 0.0056 \times 135 - 1.2 \times 0.000258\right)} \approx 0.67$$

Les deux modèles permettent de classifier les deux chansons correctement pour un seuil de 50% ; la première chanson est aimée avec une probabilité $< 50\%$ et donc Spotify ne la suggèrera pas, tandis que la deuxième chanson est aimée avec une probabilité $> 50\%$ et donc Spotify pourrait la suggerer. On ne sait pas si, en réalité, le seuil est fixé à 50% par Spotify. En plus, deux chansons ne suffisent pas pour décider de la qualité du modèle. Il faudra tester sur un grande nombre de chansons. En tout cas, on voit que les modèles ne sont pas très satisfaisants d'après la déviance : le rapport entre le `Residual deviance` et $(n - p - 1)$ est supérieur à 1 dans les deux cas.