

Données statistiques de vidéos musicales en accès sur YouTube

YouTube est un site web d'hébergement de vidéos : les internautes peuvent y déposer des vidéos, les partager et peuvent également y consulter des statistiques relatives à ces vidéos. On a choisi d'étudier certaines de ces statistiques pour n vidéos musicales, choisies parmi celles déposées en 2010 et 2011 et les plus visionnées.

On s'intéresse en particulier pour chaque vidéo au nombre de "Favoris", de "J'aime", et de "Je n'aime pas", qui correspondent respectivement aux nombres d'internautes ayant déclaré la vidéo comme une de leurs favorites, ayant déclaré l'avoir aimée, et ayant déclaré ne pas l'avoir aimée (en millions d'internautes).

On cherche ici à expliquer le nombre de "Favoris" en fonction du nombre de "J'aime" et du nombre de "Je n'aime pas".

On considère pour cela un modèle de régression linéaire multiple de la forme :

$$Y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \varepsilon_i \quad \text{pour } i \in \{1, \dots, n\},$$

où Y_i , $x_{i,1}$, $x_{i,2}$ désignent respectivement les nombres de "Favoris", de "J'aime", de "Je n'aime pas" pour la i ème vidéo considérée, et où les ε_i sont des termes d'erreur aléatoires vérifiant les conditions standards du modèle de régression linéaire multiple.

Ce modèle peut aussi s'écrire sous la forme matricielle : $Y = X\beta + \varepsilon$, avec

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \quad X = \begin{pmatrix} 1 & x_{1,1} & x_{1,2} \\ \vdots & \vdots & \vdots \\ 1 & x_{n,1} & x_{n,2} \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}, \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

On a alors

$$X'X = \begin{pmatrix} 42 & 7.763 & 0.606 \\ 7.763 & 3.843 & 0.281 \\ 0.606 & 0.281 & 0.059 \end{pmatrix}, \quad (X'X)^{-1} = \begin{pmatrix} 0.038 & -0.074 & -0.038 \\ -0.074 & 0.544 & -1.828 \\ -0.038 & -1.828 & 26.043 \end{pmatrix},$$

et si y désigne la valeur observée de Y , $X'y = (7.686, 3.682, 0.236)'$.

On suppose que le vecteur des bruits ε est de loi gaussienne.

1. Préciser l'intérêt de chacune des hypothèses du modèle. Selon vous, ces hypothèses sont-elles pertinentes pour l'étude sur les vidéos musicales de YouTube menée ici ?
2. Quel est le nombre n de vidéos considérées dans cette étude ?
3. Rappeler la définition de l'estimateur des moindres carrés ordinaires $\hat{\beta}$ de β , et donner une interprétation géométrique du vecteur des valeurs ajustées $\hat{Y} = X\hat{\beta}$. En déduire une expression matricielle de $\hat{\beta}$, puis une interprétation géométrique du vecteur des résidus $\hat{\varepsilon} = Y - \hat{Y}$.
4. Donner les lois de $\hat{\beta}$, de $\hat{Y}$, et de $\hat{\varepsilon}$. Ces vecteurs aléatoires sont-ils indépendants ?
5. La valeur de l'estimateur des moindres carrés ordinaires $\hat{\beta}$ calculé sur les observations est $\hat{\beta}_{obs} = (0.011, 1.003, -0.877)'$. Expliquer comment cette valeur a pu être obtenue à partir des données fournies ci-dessus.

6. Ecrire l'équation d'analyse de la variance, et rappeler la définition du coefficient de détermination R^2 dans le modèle considéré. Donner l'interprétation géométrique de ces deux notions.

7. Montrer que la somme des carrés expliquée est égale à $\hat{\beta}'X'X\hat{\beta} - n\bar{Y}^2$, et calculer sa valeur à partir des observations.

8. La valeur du coefficient de détermination calculé sur les observations est $R^2_{obs} = 0.981$. En déduire les valeurs de la somme des carrés totale, puis de la somme des carrés résiduelle calculées sur les observations.

9. Déterminer un estimateur sans biais de la variance du modèle. Calculer sa valeur.

10. On souhaite prédire le nombre de "Favoris" d'une vidéo qui aurait 0.4 (millions) de "J'aime" et 0.02 (millions) de "Je n'aime pas". Donner un intervalle de prédiction de niveau de confiance 95% pour ce nombre de "Favoris".

11. Tester l'hypothèse $(H_0) \beta_1 \leq 0$ contre $(H_1) \beta_1 > 0$ au niveau 5%, puis l'hypothèse $(H_0) \beta_2 \geq 0$ contre $(H_1) \beta_2 < 0$ au niveau 5%. Les résultats obtenus sont-ils surprenants ?