Université du Littoral Côte d'Opale
Probabilités & Statistiques
Ing 1
Année 2024-2025

Fiche n°6. Régression linéaire.

Avant-propos. Dans cette fiche, on considère dans chaque exercice un modèle linéaire gaussien défini par :

$$Y_i = ax_i + b + \epsilon_i,\ 1 \leq i \leq n,$$

où $(Y_1, \ldots, Y_n)$ est un n-échantillon et où les variables aléatoires ϵ_i sont indépendantes identiquement distribuées de loi $\mathcal{N}(0, \sigma^2)$. La quantité $(x_1, \ldots, x_n)$ est un n-uplet de valeurs observées et les paramètres inconnus sont a et b. On rappelle que si l'on désigne par $\hat{a}$ et $\hat{b}$ les estimateurs de a et b obtenus par la méthode des moindres carrés, alors on a les résultats suivants :

(i) les lois des estimateurs sont :

$$\hat{a} \sim \mathcal{N}\left(a, \frac{\sigma^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right) \quad \text{et} \quad \hat{b} \sim \mathcal{N}\left(b, \frac{\sigma^2 \sum_{i=1}^n x_i^2}{n \sum_{i=1}^n (x_i - \overline{x})^2}\right);$$

(ii) la loi de l'estimateur sans biais de la variance σ^2 (associée à l'hypothèse d'homoscédasticité), notée $\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n \hat{\epsilon_i}^2$, où $\hat{\epsilon}_i = y_i - (\hat{a}x_i + \hat{b})$, est telle que :

$$\frac{n-2}{\sigma^2} \hat{\sigma}^2 \sim \chi^2(n-1);$$

(iii) les variables aléatoires $(\hat{a}, \hat{b})$ et $\hat{\sigma}^2$ sont indépendantes.

Exercice 1. Un festival de jeu a lieu chaque année du côté de Lille. A la 7-ième édition, les organisateurs veulent faire une prévision du nombre de visiteurs. Pour cela, ils s'intéressent au lien entre le nombre d'internautes inscrits sur l'événement Facebook trois jours avant le festival et le nombre de visiteurs pendant le festival. Ils ont les données suivantes :

- En 2014, 141 inscrits sur Facebook pour 2 000 visiteurs ;
- En 2015, 275 inscrits sur Facebook pour 3 154 visiteurs ;
- En 2016, 344 inscrits sur Facebook pour 3 881 visiteurs ;
- En 2017, 505 inscrits sur Facebook pour 4525 visiteurs ;
- En 2018, ils ont 784 inscrits Facebook avant le festival.

On modélise le problème par une régression linéaire. Pour cela, on considère un n-échantillon $(Y_1, \ldots, Y_n)$ dont la loi est donnée par :

$$Y_i = ax_i + b + \epsilon_i,\ 1 \leq i \leq n,$$

où $(Y_1, \ldots, Y_n)$ est un n-échantillon et où les variables aléatoires ϵ_i sont indépendantes identiquement distribuées de loi $\mathcal{N}(0, \sigma^2)$.

(1) Calculer le coefficient de corrélation linéaire.

(2) Estimer a et b par la méthode des moindres carrés.

(3) Calculer les résidus $\hat{\epsilon_i} = y_i - (\hat{a}x_i + b)$.

(4) Déterminer des intervalles de confiance à 95% pour a et b.

(5) Estimer la fréquentation prévue pour cette 7-ième édition, en 2018, puis donner un intervalle de confiance à 90% pour cette estimation.