Département MIDO
Module Analyse de Données
2015 / 2016

Travaux Dirigés n° 3 : Régression linéaire simple – partie II

Objectifs : revoir des notions vues en cours. Savoir calculer les propriétés des estimateurs des MCO.

1 Partie 1 : questions de cours

1.1 Exercice 1 - Erreur de prévision

Montrez que l'erreur de prévision $\hat{\varepsilon}_{n+1} = (y_{n+1} - \hat{y}_{n+1})$ satisfait les propriétés suivantes :

$$\begin{cases} \mathbb{E}(\hat{\varepsilon}_{n+1}) & = 0 \\ \text{var}(\hat{\varepsilon}_{n+1}) & = \sigma^2 \left(1 + \frac{1}{n} + \frac{(x_{n+1} - \overline{\mathbf{x}})^2}{\sum_{i=1}^{n}(x_i - \overline{\mathbf{x}})^2}\right) \end{cases} \tag{1}$$

1.2 Exercice 2 : Décomposition de la variance et coefficient de détermination

D'après le cours, nous avons :

$$||\mathbf{y} - \overline{\mathbf{y}}\mathbb{1}||^2 = ||\hat{\mathbf{y}} - \overline{\mathbf{y}}\mathbb{1}||^2 + ||\hat{\varepsilon}||^2 \tag{2}$$

$$\sum_{i=1}^{n}(y_i - \overline{\mathbf{y}})^2 = \sum_{i=1}^{n}(\hat{y}_i - \overline{\mathbf{y}})^2 + \sum_{i=1}^{n}\hat{\varepsilon}_i^2, \tag{3}$$

$$\text{SCT} = \text{SCE} + \text{SCR} \tag{4}$$

- SCT : *Somme des Carrés des écarts Totale* ; elle possède $(n-1)$ degrés de liberté
- SCE : *Somme des Carrés des écarts Expliquée* ; elle possède (1) degrés de liberté
- SCR : *Somme des Carrés des écarts Résiduelle* ; elle possède $(n-2)$ degrés de liberté

Montrez que :

$$R^2 = \frac{\text{SCE}}{\text{SCT}} = \frac{||\hat{\mathbf{y}} - \overline{\mathbf{y}}\mathbb{1}||^2}{||\mathbf{y} - \overline{\mathbf{y}}\mathbb{1}||^2} = 1 - \frac{||\hat{\varepsilon}||^2}{||\mathbf{y} - \overline{\mathbf{y}}\mathbb{1}||^2} = 1 - \frac{\text{SCR}}{\text{SCT}} = \rho^2_{\mathbf{xy}}$$

Expliquer en vous appuyant sur une représentation géométrique (cf. cours) ce qui se passe quand :

- $R^2 = 1$
- $R^2 = 0$
- R^2 est proche de zéro

2 Partie 2 : pratique de la régression

On appelle "fréquence seuil" d'un sportif amateur sa fréquence cardiaque obtenue après trois quarts d'heure d'un effort soutenu de course à pied. Celle-ci est mesurée à l'aide d'un cardio-fréquence-mètre. On cherche à savoir si l'âge d'un sportif a une influence sur sa fréquence seuil. On dispose pour cela de 20 valeurs du couple (x_i, y_i), où x_i est l'âge et y_i la fréquence seuil du sportif. On a obtenu $(\overline{\mathbf{x}}, \overline{\mathbf{y}}) = (35,6; 170,2)$ et :

$$\sum_{i=1}^{n}(x_i - \overline{\mathbf{x}})^2 = 1991, \ \sum_{i=1}^{n}(y_i - \overline{\mathbf{y}})^2 = 189,2, \ \sum_{i=1}^{n}(x_i - \overline{\mathbf{x}})(y_i - \overline{\mathbf{y}}) = -195,4$$

1. Calculer la droite des moindres carrés pour le modèle $y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$.
2. Avec ces estimateurs, la somme des carrés des résidus vaut $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = 170$. Si on suppose les perturbations ε_i sont de même variance σ^2 et d'espérance nulle, en déduire un estimateur non biaisé $\hat{\sigma}^2$ de σ^2.
3. Calculer un estimateur de la variance de $\hat{\beta}_2$.
4. Calculer les variances SCE, SCT et SCR et en déduire le coefficient de détermination. Commenter la qualité de l'ajustement des données au modèle.