UNIVERSITÉ CÔTE D'AZUR

Cédric Richard

Méthodes de régression

Exercice 1
Soit $\{(y_i, x_i)\}_{i=1}^n$ un ensemble donné de points de $\mathbb{R}^2$. On appelle estimateurs de Moindres Carrés Ordinaires (OLS) les variables $\widehat{\beta}_1$ et $\widehat{\beta}_2$ minimisant le critère :

$$J(\beta_1, \beta_2) = \sum_{i=1}^{n} (y_i - \beta_1 - \beta_2 \, x_i)^2$$

La résolution de celui-ci vise à estimer la droite des moindres carrés d'équation $y = \beta_1 + \beta_2 \, x$.

1. Montrer que $J(\beta_1, \beta_2)$ admet un minimum unique $(\widehat{\beta}_1, \widehat{\beta}_2)$.
2. Poser les conditions d'optimalité.
3. En déduire que $\widehat{\beta}_1 = \bar{y} - \widehat{\beta}_2 \bar{x}$, où l'on précisera les expressions et la nature de $\bar{y}$ et $\bar{x}$. Interpréter ce résultat.
4. De même, montrer que :
$$\widehat{\beta}_2 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$
5. En déduire l'expression alternative :
$$\widehat{\beta}_2 = \frac{\sum_{i=1}^n (x_i - \bar{x}) \, y_i}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

Les expressions de $\widehat{\beta}_1$ et $\widehat{\beta}_2$ ci-dessus pourront être utilisées dans la suite.

Exercice 2
L'étude statistique ci-dessous porte sur les poids respectifs des pères et de leur fils aîné.

Père (p_i)	65	63	67	64	68	62	70	66	68	67	69	71
Fils (f_i)	68	66	68	65	69	66	68	65	71	67	68	70

On rapporte les résultats numériques suivants, obtenus à partir des données ci-dessus.

$$\sum_{i=1}^{12} p_i = 800 \quad \sum_{i=1}^{12} p_i^2 = 53418 \quad \sum_{i=1}^{12} p_i f_i = 54107 \quad \sum_{i=1}^{12} f_i = 811 \quad \sum_{i=1}^{12} f_i^2 = 54849$$

1. Estimer la droite des moindres carrés du poids des fils en fonction de celui des pères.
2. Estimer la droite des moindres carrés du poids des pères en fonction de celui des fils.
3. Montrer que le produit des pentes de ces 2 droites est égal au carré du coefficient de corrélation empirique entre les p_i et les f_i.

Exercice 3
Une étude a visé à exprimer la hauteur y en pieds d'arbres d'une essence donnée en fonction de leur diamètre x en pouces, à 1 mètre 30 du sol. Pour cela, 20 couples de mesures (x_i, y_i) ont été effectués et on permit d'aboutir aux informations suivantes :

$$\bar{x} \triangleq \frac{1}{20} \sum_{i=1}^{20} x_i = 4.53 \quad \bar{y} \triangleq \frac{1}{20} \sum_{i=1}^{20} y_i = 8.65$$

$$\frac{1}{20} \sum_{i=1}^{20} (x_i - \bar{x})^2 = 10.97 \quad \frac{1}{20} \sum_{i=1}^{20} (y_i - \bar{y})^2 = 2.24 \quad \frac{1}{20} \sum_{i=1}^{20} (x_i - \bar{x})(y_i - \bar{y}) = 3.77$$

1. Calculer les coefficients $\widehat{\beta}_1$ et $\widehat{\beta}_2$ de la droite de régression.
2. Donner et commenter une mesure de qualité de l'ajustement du modèle aux données.
3. Exprimer cette mesure en fonction des statistiques élémentaires. Commenter le résultat.

Exercice 4
Douze personnes sont inscrites à une formation. Au début de la formation, ces stagiaires subissent une épreuve A notée sur 20. A la fin de la formation, ils subissent à nouveau une épreuve B notée sur 20 d'un niveau identique afin d'évaluer leurs progrès. Les résultats sont donnés dans le tableau suivant :

Epreuve A	3	4	6	7	9	10	9	11	12	13	15	4
Epreuve B	8	9	10	13	15	14	13	16	13	19	6	19

1. Représenter le nuage de points. Déterminer la droite de régression. Calculer le coefficient de détermination. Commenter.
2. Deux stagiaires se distinguent des autres. Les supprimer et déterminer la droite de régression sur les points restants. Calculer le coefficient de détermination. Commenter.

Exercice 5
Une étude s'est intéressée à la hauteur y en mètres des eucalyptus en fonction de leur circonférence x en centimètres, à 1 mètre 30 du sol. Pour ce faire, $n = 1429$ mesures (x_i, y_i) ont été effectuées et on permit d'aboutir aux informations suivantes :

$$\bar{x} \triangleq \frac{1}{n}\sum_{i=1}^{n} x_i = 47.3 \quad \bar{y} \triangleq \frac{1}{n}\sum_{i=1}^{n} y_i = 21.2$$

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = 102924 \quad \sum_{i=1}^{n}(y_i - \bar{y})^2 = 8857 \quad \sum_{i=1}^{20}(x_i - \bar{x})(y_i - \bar{y}) = 26466$$

1. Estimer la droite des moindres carrés pour le modèle $y = \beta_1 + \beta_2 x + \varepsilon$, et la représenter sur la Figure 1.
2. Calculer le coefficient de détermination. Commenter la qualité de l'ajustement des données au modèle.
3. Avec ces estimateurs, la somme des carrés des résidus vaut $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = 2052$. Si on suppose les perturbations ε_i gaussiennes, centrées, indépendantes, de même variance σ^2, en déduire un estimateur non biaisé $\hat{\sigma}^2$ de σ^2.
4. Donner un estimateur $\hat{\sigma}_1^2$ de la variance de $\hat{\beta}_1$.
5. Une analyse de la Figure 1 montre que, pour les petites circonférences et hauteurs, un comportement en racine carrée pourrait représenter les données de manière appropriée. Ecrire le modèle de régression correspondant. Ecrire le problème de régression associé sous forme matricielle.

[1] Une partie de ces exercices est extraite du premier chapitre du cours d'Arnaux Guyader, UPMC.

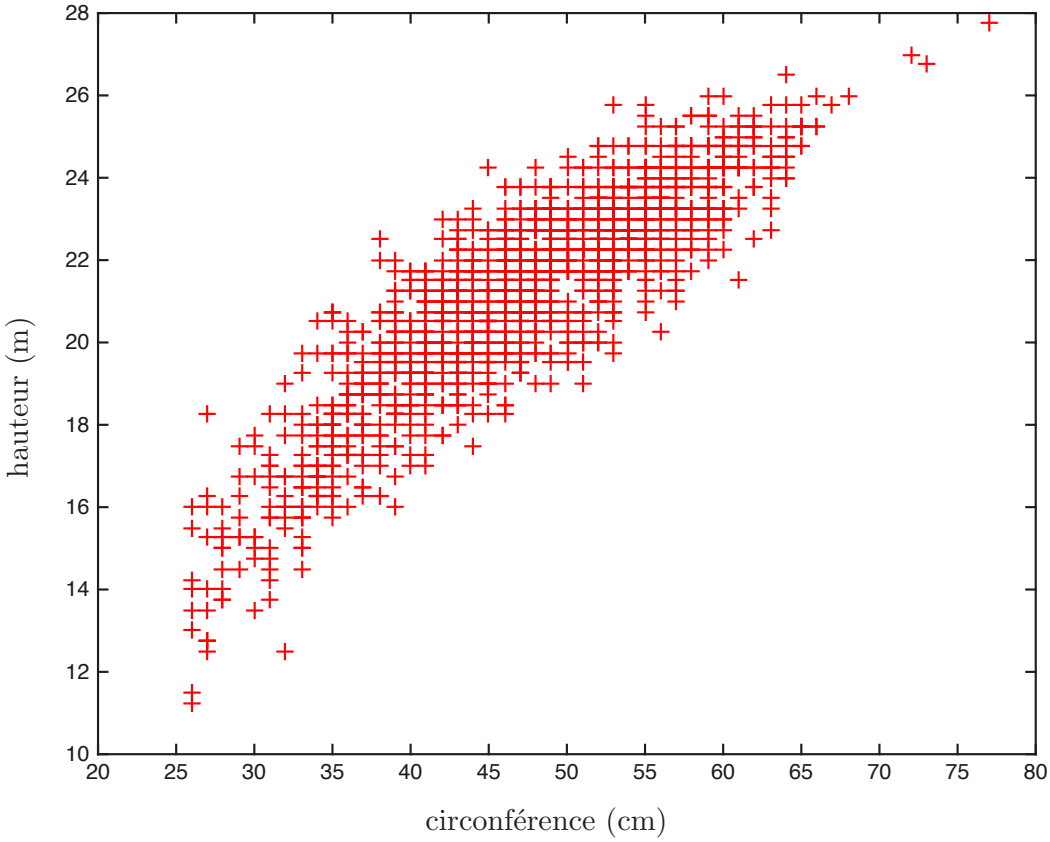

Figure 1: Eucalyptus