TD DE MODÈLE LINÉAIRE - FEUILLE 3
Régression linéaire simple

Exercice 1 Rappeler la formule définissant le coefficient de détermination R^2 et la développer pour montrer qu'il est égal au carré du coefficient de corrélation empiriqueentre x et y, noté $\rho_{x,y}$, c'est-à-dire :

$$R^2 = \rho_{x,y}^2 = \left(\frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} \right)^2 .$$

Exercice 2 Considérons le modèle statistique suivant :

$$y_i = \beta x_i + \varepsilon_i, \qquad i = 1, \cdots, n,$$

où nous supposons que les perturbations ε_i sont telles que $\mathbb{E}[\varepsilon_i] = 0$ et $\mathrm{Cov}(\varepsilon_i, \varepsilon_i) = \sigma^2 \delta_{i,j}$.

1. En revenant à la définition des moindres carrés, montrer que l'estimateur des moindres carrés de β vaut $\hat{\beta} = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}$.
2. Montrer que la droite passant par l'origine et le centre de gravité du nuage de points est $y = \beta^* x$, avec $\beta^* = \frac{\sum_{i=1}^n y_i}{\sum_{i=1}^n x_i}$.
3. Montrer que $\hat{\beta}$ et β^* sont des estimateurs sans biais de β.
4. Montrer que $V(\beta^*) > V(\hat{\beta})$ sauf dans le cas où tous les x_i sont égaux (penser à l'inégalité de Cauchy-Schwarz). Ce résultat était-il prévisible ?

Exercice 3 On appelle "fréquence seuil" d'un sportif amateur sa fréquence cardiaque obtenue aprés 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 $(\bar{x}, \bar{y}) = (35,6; 170,2)$ et :

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

1. Calculer les estimateurs des moindres carrés pour le modèle $y = \beta_1 + \beta_2 x + \varepsilon$.
2. Question de cours : montrer que

$$\hat{\beta}_2 = \beta_2 + \frac{\sum (x_i - \overline{x})\varepsilon_i}{\sum (x_i - \overline{x})^2}.$$

3. Calculer le coefficient de détermination R^2. Commenter la qualité de l'ajustement des données au modèle.
4. Avec ces estimateurs, la somme des carrés des résidus vaut $\sum_{i=1}^n (y_i - \hat{y}_i)^2 = 170$. On suppose les perturbations ε_i gaussiennes, indépendantes et de même variance σ^2. Déduire un estimateur non biaisé $\hat{\sigma}^2$ de σ^2
5. Déduire un estimateur $\hat{\sigma}_2^2$ de la variance de $\hat{\beta}_2$.

6. Calculer l'intervalle de confiance au risque 5% de β_2.
7. Tester l'hypothèse $H_0 : \beta_2 = 0$ contre $H_1 : \beta_2 \neq 0$ pour un risque de 5%. Conclure sur la question de l'influence de l'âge sur la fréquence seuil.

Exercice 4 On considère le modèle de régression linéaire simple $y = \beta_1 + \beta_2 x + \epsilon$. Soit un échantillon $(x_i, y_i)_{1 \leq i \leq 100}$ de statistiques résumées

$$\sum_{i=1}^{100} x_i = 0 \qquad \sum_{i=1}^{100} x_i^2 = 400 \qquad \sum_{i=1}^{100} x_i y_i = 100 \qquad \sum_{i=1}^{100} y_i = 100 \qquad \hat{\sigma}^2 = 1$$

1. Exprimer les intervalles de confiance à 95% pour β_1 et β_2.
2. Donner l'équation de la région de confiance à 95% de (β_1, β_2). Rappel : $\frac{(x-x_0)^2}{a^2} + \frac{(y-y_0)^2}{b^2} \leq 1$ est l'équation de l'intérieur de l'ellipse centré en (x_0, y_0), dont les axes sont parallèles à ceux des abscisses et des ordonnées, et de sommets $(x_0 \pm a, y_0)$ et $(x_0, y_0 \pm b)$.
3. Représenter sur un même graphique les résultats obtenus.

Exercice 5 Au $17^{\text{ème}}$ siècle, Huygens s'est intéressé aux forces de résistance d'un objet en mouvement dans un fluide (eau, air, etc.). Il a d'abord émis l'hypothèse selon laquelle les forces de frottement étaient proportionnelles à la vitesse de l'objet, puis, après expérimentation, selon laquelle elles étaient proportionnelles au carré de la vitesse. On réalise une expérience dans laquelle on fait varier la vitesse x d'un objet et on mesure les forces de frottement y.

1. Quel(s) modèle(s) testeriez-vous ?
2. Comment feriez-vous pour déterminer le modèle adapté ?

Exercice 6 On souhaite expliquer la hauteur y (en mètres) d'un arbre en fonction de sa circonférence x (en centimètres) à 1m30 du sol. On a relevé $n = 1429$ couples (x_i, y_i). On a obtenu $(\bar{x}, \bar{y}) = (47,3; 21,2)$ et :

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

1. Calculer la droite des moindres carrés pour le modèle $y = \beta_1 + \beta_2 x + \epsilon$.
2. Calculer le coefficient de détermination R^2. Commenter la qualité de l'ajustement des données au modèle.
3. Avec ces estimateurs, la somme des carrés des résidus vaut alors $\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = 2052$. Si on suppose les perturbations ϵ_i gaussiennes, centrées, indépendantes et 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. Tester l'hypothèse $H_0 : \beta_1 = 0$ contre $H_1 : \beta_1 \neq 0$.