TD N° 5 : Modèle linéaire

Exercice 1. (Prédiction) On prend l'exemple d'un nouveau médicament contre l'allergie, et l'on étudie le réglage de son dosage. Les données sont les suivantes (pour les 10 patients de l'expérience) :

dosage (mg)	3	3	4	5	6	6	7	8	8	9
soulagement (jours)	9	5	12	9	14	16	22	18	24	22

1) Représenter le nuage de points (x_i, y_i) pour $i = 1, \dots, 10$.
2) On donne les quantités $\sum_{i=1}^{10} x_i y_i = 1003$, $\sum_{i=1}^{10} x_i^2 = 389$, $\sum_{i=1}^{10} y_i^2 = 2651$, $\bar{x}_n = 5.9$ et $\bar{y}_n = 15.1$. Déterminer les coefficients estimés de la droite de régression des moindres carrés $\hat{\beta}_0$ et $\hat{\beta}_1$. Représenter la droite des moindres carrés sur le graphique précédent.
3) Déterminer un intervalle de confiance de niveau 95% pour β_0 et β_1.
4) Donner une prédiction $\hat{y}$ du nombre de jours de soulagement pour un dosage $x^* = 4.5\,mg$. Déterminer un intervalle de confiance de cette prédiction de niveau 95%.
5) Déterminer le pourcentage de variabilité expliqué par le modèle (coefficient R^2). Interpréter.

Exercice 2. Utilisez les informations contenues dans le Tableau 1 pour :

1) Trouver la droite des moindres carrés prédisant la taille avant mue en fonction de la taille après mue.
Aide : montrer que
$$\hat{\beta}_1 = R\frac{s_n(\mathbf{y})}{s_n(\mathbf{x})} \quad . \tag{1}$$
où R est le coefficient de corrélation de Spearman que l'on peut définir par :
$$R = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{\mathrm{cov}_n(\mathbf{x}, \mathbf{y})}{s_n(\mathbf{x})s_n(\mathbf{y})}$$

Table 1 – Statistiques résumées de la taille des carapaces de crabes dormeurs femelles.

	Moyenne	Écart-type
Avant mue	129mm	11mm
Après mue	144mm	10mm

$R = 0.98$

2) Trouvez la droite des moindres carrés prédisant l'accroissement de la taille en fonction de la taille après mue sachant que la corrélation entre la taille après mue et l'accroissement est -0.45 et que l'accroissement moyen est de 14.7mm avec un écart-type de 2.2mm.

Exercice 3. Le Tableau 2 contient les moyennes, écart-types pour la masse des souris mâles et femelles issues de la famille 141G6. On se propose d'étudier le modèle $\mathbb{E}[y_i] = \beta_0^\star + \beta_1^\star \mathbb{1}_{M,i}$, où $\mathbb{1}_{M,i}$ est une variable binaire valant 1 si la i-ième souris est un mâle et 0 sinon.

Table 2 – Statistiques sur la masse (g) des souris issues de la famille 141G6 (transgéniques ou non).

	Nombre	Moyenne	Écart-type
Mâle	94	31.70	2.62
Femelle	83	25.23	2.00

1) Que représentent $\beta_0^\star$ et $\beta_1^\star$ dans ce modèle ?

2) Utiliser les statistiques résumées du Tableau 2 pour estimer les coefficients du modèle

3) On notera n_M (resp n_F) le nombre de mâles (resp. de femelles) et $\hat{\sigma}_M$ (resp. $\hat{\sigma}_F$) les estimateurs sans biais de l'écart-type des mâles (resp. des femelles). On veut calculer l'estimateur de la variance de $\hat{\beta}_1$ vu en cours ; rappelons que celui-ci est donné en général par :

$$\hat{\sigma}_1^2 := \frac{\hat{\sigma}^2}{n} \frac{1}{\operatorname{var}_n(\mathbf{x})} \ ,$$

avec $\mathbf{x} = \mathbb{1}_M \in \mathbb{R}^n$.
Montrer que ici, on a :

$$n \operatorname{var}_n(\mathbf{x}) = \frac{n_M n_F}{n_M + n_F}.$$

et

$$\hat{\sigma}^2 = \frac{(n_M - 1)\hat{\sigma}_M^2 + (n_F - 1)\hat{\sigma}_F^2}{n_M + n_F - 2}.$$

En déduire que :

$$\hat{\sigma}_1 \simeq 0.35.$$

4) Faire un test (bilatéral) pour l'hypothèse $\mathcal{H}_0$: "$\beta_1^\star = 0$". On donne 1.97 la valeur du quantile d'ordre 0.975 d'une loi de Student à 175 degrés de liberté.