## TD n° 5 : Modèle linéaire

**Exercice 1. (QQ-plot (retour du TD1))** ⚠ Exercice à faire sur machine.

Ci-dessous vous (re-)trouverez les quantiles associés aux probabilités $0.05, 0.10, \ldots, 0.95$ de la durée de la grossesse (en jours) pour les mères de l'étude CHDS vue en cours. Tracez (idéalement utiliser un notebook pour cela) ces quantiles en fonction de ceux d'une loi $\mathcal{U}(0,1)$ uniforme sur le segment $[0,1]$. Décrire la forme de la distribution de la durée de la grossesse par rapport à la loi uniforme. Utilisez un modèle de régression linéaire pour obtenir la pente et l'ordonnée à l'origine d'une approximation linéaire de ce graphique. Interpretez les valeurs obtenues.

250, 262, 267, 270, 272, 274, 275, 277, 278, 280, 282, 283, 284, 286, 288, 290, 292, 295, 302.

**Exercice 2. (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 $i = 1, \ldots, 10$.
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 $\beta_0$ et $\beta_1$.
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.

**Correction:**

Rappel :

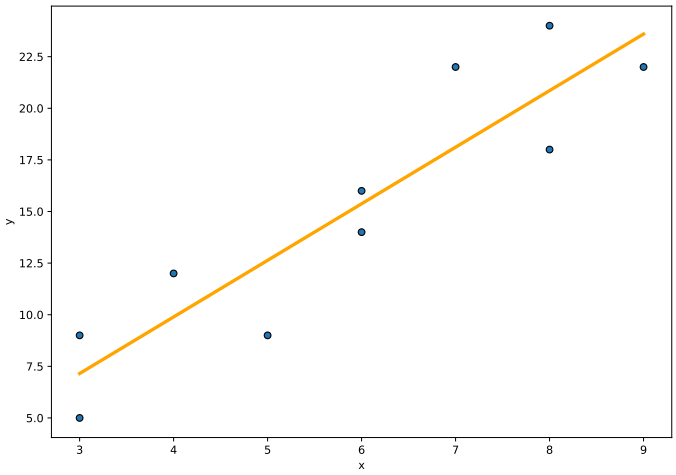
$$S_{xx} = \sum_i (x_i - \bar{x})^2 = n \operatorname{var}_n(\mathbf{x}) = \sum_i x_i^2 - \frac{1}{n}\left(\sum_i x_i\right)^2$$

$$S_{yy} = \sum_i (y_i - \bar{y})^2 = n \operatorname{var}_n(\mathbf{y}) = \sum_i y_i^2 - \frac{1}{n}\left(\sum_i y_i\right)^2$$

$$S_{xy} = \sum_i (x_i - \bar{x})(y_i - \bar{y}) = n \operatorname{cov}_n(\mathbf{x}, , \mathbf{y}) = \sum_i x_i y_i - \frac{1}{n}\left(\sum_i x_i \sum_i y_i\right)$$

$$SSE = \sum_i (y_i - \hat{y}_i)^2 = S_{yy} - \frac{S_{xy}^2}{S_{xx}}.$$

1) Nuage de points avec droite de régression :




2) $\hat{\beta}_1 = S_{xy}/S_{xx}$ Or :

$$S_{xy} = 1003 - \frac{1}{10}\left(\sum_i x_i \sum_i y_i\right) = 1003 - \frac{59 \times 151}{10} = 112.1$$

$$S_{xx} = 389 - \frac{1}{10}\left(\sum_i x_i\right)^2 = 389 - \frac{3481}{10} = 40.9$$

Donc $\hat{\beta}_1 = 112.1/40.9 \simeq 2.74$ et $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 15.1 - 2.74 \times 5.9 \simeq -1.07$

3) Au niveau $(1-\alpha)$, l'I.C. de $\beta_1$ est :

$$\left[\hat{\beta}_1 - t_{1-\alpha/2}(8)\frac{S}{\sqrt{S_{xx}}} \; ; \hat{\beta}_1 + t_{1-\alpha/2}(8)\frac{S}{\sqrt{S_{xx}}}\right] \quad \text{Où } S = \sqrt{\hat{\sigma}^2} = \sqrt{\frac{SSE}{n-2}}$$

Ici :

$$\begin{aligned} SSE &= S_{yy} - \frac{S_{xy}^2}{S_{xx}} \\ &= \sum_i y_i^2 - \frac{1}{n}\left(\sum_i y_i\right)^2 - \frac{S_{xy}^2}{S_{xx}} = 2651 - \frac{151^2}{10} - \frac{112.1^2}{40.9} \simeq 63.65 \; . \end{aligned}$$

On en déduit :

$$\begin{aligned} S &= \sqrt{\frac{63.65}{8}} \simeq 2.82 \\ \hat{\beta}_1 - t_{1-\alpha/2}(8)\frac{S}{\sqrt{S_{xx}}} &\simeq 2.74 - t_{0.975}(8)\frac{2.82}{\sqrt{40.9}} \\ &\simeq 2.74 - 2.306 \times \frac{2.82}{6.395} \\ &\simeq 1.72 \\ \hat{\beta}_1 + t_{1-\alpha/2}(8)\frac{S}{\sqrt{S_{xx}}} &\simeq 2.74 + t_{0.975}(8)\frac{2.82}{\sqrt{40.9}} \\ &\simeq 3.76 \end{aligned}$$

Donc l'intervalle de confiance au niveau 95% est $[1.72; 3.76]$ (i.e. avec une confiance de 95%, en ajoutant 1 mg de médicament, on ajoute en moyenne entre 1.72 et 3.76 jours de soulagement.)

Au niveau $(1-\alpha)$, l'I.C. de $\beta_0$ est :

$$\left[\hat{\beta}_0 - t_{1-\alpha/2}(8)\, S\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}}\,;\hat{\beta}_0 + t_{1-\alpha/2}(8)\, S\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}}\right]$$

Ici :

$$\begin{aligned} S &\simeq 2.82 \\ \hat{\beta}_0 - t_{1-\alpha/2}(8)\, S\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}} &\simeq -1.07 - t_{0.975}(8)\, S\sqrt{\frac{1}{10}+\frac{34.81}{40.9}} \\ &\simeq -1.07 - (2.306)\times(2.82)\times\sqrt{0.951} \\ &\simeq -7.41 \\ \hat{\beta}_0 + t_{1-\alpha/2}(8)\, S\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}} &\simeq -1.07 + t_{0.975}(8)\, S\sqrt{\frac{1}{10}+\frac{34.81}{40.9}} \\ &\simeq 5.27 \end{aligned}$$

4) Prédiction :

$$\begin{aligned} y^* &= \hat{\beta}_0 + x^*\hat{\beta}_1 \\ &\simeq -1.07 + 4.5\times 2.74 \\ &\simeq 11.26 \end{aligned}$$

Et l'I.C. de $\beta_0+\beta_1 x^*$ au niveau $1-\alpha$ est :

$$\left[\ \hat{\beta}_0+\hat{\beta}_1 x^* \pm t_{1-\alpha/2}(n-2)\, S\sqrt{\frac{1}{n}+\frac{(x^*-\bar{x})^2}{S_{xx}}}\ \right]$$

Donc ici, l'IC de $y^*$ à 95% est $[8.76\,;13.76]$.

5) La **proportion de la variabilité** de $Y$ **expliquée** par le modèle est

$$r^2 = \frac{S_{xy}^2}{S_{xx}S_{yy}}.$$

Ici : $r^2 = \frac{112.1^2}{40.9\times 370.9} \simeq 0.828 = 82.8\%$

**Exercice 3.** 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

$$\hat{\beta}_1 = R\frac{s_n(\mathbf{y})}{s_n(\mathbf{x})}\ . \tag{1}$$

où $R$ est le coefficient de corrélation de Spearman que l'on peut définir par :

$$R = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{\mathrm{cov}_n(\mathbf{x},\mathbf{y})}{s_n(\mathbf{x})s_n(\mathbf{y})}$$

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.

**Correction:**

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

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

$R = 0.98$

1) Pour montrer la relation demandée, on commence par vérifier une égalité préliminaire. Puisque

$$S_{xx} = n \operatorname{var}_n(\mathbf{x}) \text{ et } S_{yy} = n \operatorname{var}_n(\mathbf{y}),$$

on a :

$$\sqrt{\frac{S_{yy}}{S_{xx}}} = \frac{s_n(\mathbf{y})}{s_n(\mathbf{x})}.$$

On peut maintenant montrer l'égalité demandée :

$$R \frac{s_n(\mathbf{y})}{s_n(\mathbf{x})} = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}} \frac{\sqrt{S_{yy}}}{\sqrt{S_{xx}}} = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{xx}}} = \frac{S_{xy}}{S_{xx}} = \hat{\beta}_1$$

**Attention ! !** ici on veut prédire la taille avant mue en fonction de la taille après mue, donc le $\mathbf{x}$ correspond à **après** mue et le $\mathbf{y}$ à **avant** mue.
Donc ici :

$$\hat{\beta}_1 = 0.98 \times \frac{11}{10} \simeq 1.078$$

et

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 129 - 1.078 \times 144 \simeq -26.23.$$

2) La corrélation entre la taille après mue et l'accroissement est $R = -0.45$, donc on en déduit que :

$$\hat{\beta}_1 = R \frac{s_n(\mathbf{y})}{s_n(\mathbf{x})} = -0.45 \times \frac{2.2}{10} \approx -0.099 \ ,$$

et

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 14.7 + 0.099 \times 144 \approx 28.956 \, .$$

**EXERCICE 4.** 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.

**Correction:**

1) Dans le modèle proposé, $\beta_0^\star$ représente la masse moyenne d'un échantillon infini lorsque $\mathbb{1}_M$ vaut 0, autrement dit lorsque toutes les souris sont de sexe femelle.
$\beta_0^\star + \beta_1^\star$ représente la masse moyenne d'un échantillon infini lorsque $\mathbb{1}_M$ vaut 1, autrement dit lorsque toutes les souris sont des mâles. Ainsi, $\beta_1^\star$ représente la différence entre la masse moyenne des mâles et celle des femelles lorsque l'échantillon est infini.

2) $\beta_0^\star$ est estimé sur l'échantillon de taille finie décrit dans la table ci-dessus par

$$\hat{\beta}_0 = \bar{y}_F = 25.23.$$

Et $\beta_1$ est estimé par

$$\hat{\beta}_1 = \bar{y}_M - \bar{y}_F = 31.70 - 25.23 = 6.47.$$

3) Par définition on a :

$$n \operatorname{var}_n(x) = \sum_i (x_i - \bar{x})^2,$$

où $n$ est le nombre total d'individus. Dans la somme ci-dessus, on doit distinguer deux cas : si l'individu est un mâle alors $x_i = 1$ et si l'individu est une femelle alors $x_i = 0$ ce qui permet d'écrire :

$$\begin{aligned} n \operatorname{var}_n(x) = \sum_i (x_i - \bar{x})^2 &= \sum_{i:\text{mâles}} (1 - \bar{x})^2 + \sum_{i:\text{femelles}} (-\bar{x})^2 \\ &= n_M (1 - \bar{x})^2 + n_F \bar{x}^2 \end{aligned}$$

En utilisant la définition des $x_i$ on obtient :

$$\bar{x} = \frac{n_M}{n_M + n_F}$$

et

$$1 - \bar{x} = \frac{n_F}{n_M + n_F}.$$

On peut maintenant terminer le calcul :

$$\begin{aligned} n \operatorname{var}_n(x) = n_M (1 - \bar{x})^2 + n_F \bar{x}^2 &= n_M \left( \frac{n_F}{n_M + n_F} \right)^2 + n_F \left( \frac{n_M}{n_M + n_F} \right)^2 \\ &= \frac{n_M n_F^2 + n_F n_M^2}{(n_M + n_F)^2} \\ &= \frac{n_M n_F (n_F + n_M)}{(n_M + n_F)^2} \\ &= \frac{n_M n_F}{n_M + n_F} \\ &= \frac{1}{\frac{1}{n_M} + \frac{1}{n_F}} \end{aligned}$$

Calculons maintenant le deuxième terme demandé, celui-ci est donné par la formule générale suivante :

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}\Big[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)\Big]^2 .$$

Comme précédemment la somme ci-dessus se coupe en deux sous-sommes suivant que les individus soient des mâles ou femelles. On obtient donc :

$$\hat{\sigma}^2 = \frac{1}{n_M + n_F - 2}\left(\sum_{i:\text{mâles}} (y_i - (\hat{\beta}_0 + \hat{\beta}_1))^2 + \sum_{i:\text{femelles}} (y_i - \hat{\beta}_0)^2\right).$$

Comme $y_i$ est le poids de l'individu, d'après la question précédente $\hat{\beta}_0 = \bar{y}_F$ est la masse moyenne d'une femelle et $\hat{\beta}_0 + \hat{\beta}_1 = \bar{y}_M$ est la masse moyenne d'un mâle sur l'échantillon. Et donc :

$$\hat{\sigma}^2 = \frac{1}{n_M + n_F - 2}\left(\sum_{i:\text{mâles}} (y_i - \bar{y}_M)^2 + \sum_{i:\text{femelles}} (y_i - \bar{y}_F)^2\right).$$

Rappelons maintenant que par définition de $\hat{\sigma}_F^2$ et de $\hat{\sigma}_M^2$, on a :

$$(n_M - 1)\hat{\sigma}_M^2 = \sum_{i:\text{mâles}} (y_i - \bar{y}_M)^2$$

et

$$(n_F - 1)\hat{\sigma}_F^2 = \sum_{i:\text{femelles}} (y_i - \bar{y}_F)^2.$$

Et donc finalement on obtient :

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

4) On met en place un test de Student

$$\mathcal{H}_0 \ : \text{“}\beta_1^\star = 0\text{”} \quad vs. \quad \mathcal{H}_1 \ : \text{“}\beta_1^\star \neq 0\text{”}.$$

La statistique de test qui nous intéresse est donnée par

$$T = \frac{\hat{\beta}_1}{\hat{\sigma}_1}.$$

Calculons la réalisation $T_{\text{obs}}$ de $T$ obtenue grâce à l'échantillon. On a calculé à la question précédente $\hat{\sigma}_1 \approx 0.35$ et donc

$$T_{\text{obs}} \approx \frac{6.47}{0.35} \approx 18.49.$$

Sous $\mathcal{H}_0$, la statistique de test suit une loi de Student à $177 - 2 = 175$ degrés de liberté. Par contraste, sous l'hypothèse alternative, $T$ est très éloigné de 0 (dans les positifs ou les négatifs car le test est bilatéral), dont $|T|$ est très grand. Au risque 5%, la forme de la zone de rejet est de la forme

$$]-\infty\ ; -a] \cup [a\ ;\ +\infty[$$

où $a$ est la valeur critique égale au quantile d'ordre 0.975 pour la loi de Student à 175 degrés de libertés. Celui-ci nous est donné, il est égal à 1.97. Conclusion ? La valeur observée tombe dans la zone de rejet, donc on rejette $\mathcal{H}_0$ et la différence observée entre la masse des mâles et celle des femelles est significative. En fait, la $p$-value est ici très faible (plus petite que $10^{-22}$) donc $\beta_1^\star$ est sans doute très significativement différent de 0, ce qui permet de valider une différence entre mâle et femelle sur ce critère.