

Correction du TD n° 6

Exercice 1 :

1. Voir polycopié du cours.
2. D'une part,

$$\begin{aligned}\sum_{i=1}^n (x_i - \bar{x})^2 &= \sum_{i=1}^n (x_i^2 - 2\bar{x}x_i + \bar{x}^2) \\ &= \sum_{i=1}^n (x_i^2) - 2\bar{x} \sum_{i=1}^n (x_i) + n\bar{x}^2 \\ &= \sum_{i=1}^n (x_i^2) - n\bar{x}^2\end{aligned}$$

D'autre part,

$$\begin{aligned}\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) &= \sum_{i=1}^n (x_i y_i - \bar{x}y_i - x_i \bar{y} + \bar{x}\bar{y}) \\ &= \sum_{i=1}^n (x_i y_i) - \bar{x} \sum_{i=1}^n (y_i) - \bar{y} \sum_{i=1}^n (x_i) + n\bar{x}\bar{y} \\ &= \sum_{i=1}^n (x_i y_i) - n\bar{x}\bar{y}\end{aligned}$$

3. Exceptionnellement, on note d le coefficient de détermination. Pour rappel,

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}, \quad d = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}.$$

avec $\hat{y}_i = \hat{a}x_i + \hat{b}$. On veut montrer que $d = r^2$. On commence par

remarquer que

$$\begin{aligned}\sum_{i=1}^n (\hat{y}_i - \bar{y})^2 &= \sum_{i=1}^n (\hat{a}x_i + \hat{b} - \bar{y})^2 \\ &= \sum_{i=1}^n (\hat{a}x_i - \hat{a}\bar{x})^2 \\ &= \hat{a}^2 \sum_{i=1}^n (x_i - \bar{x})^2\end{aligned}$$

En utilisant l'expression de $\hat{a}$ donné dans la question 2, on en déduit

$$\frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = \frac{(\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}))^2}{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}$$

Autrement dit, $d = r^2$.

4. Pour montrer cette égalité, on commence par une manipulation classique sur le terme de gauche pour faire apparaître les deux termes de droites.

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2 \quad (1)$$

$$= \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + 2 \sum_{i=1}^n (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) \quad (2)$$

Il reste donc à montrer que

$$\sum_{i=1}^n (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = 0.$$

On a

$$\begin{aligned}
\sum_{i=1}^n (y_i - \hat{y}_i) (\hat{y}_i - \bar{y}) &= \hat{a} \sum_{i=1}^n (y_i - \hat{y}_i) (x_i - \bar{x}) \\
&= \hat{a} \sum_{i=1}^n (y_i - \bar{y} - \hat{a}(x_i - \bar{x})) (x_i - \bar{x}) \\
&= \hat{a} \sum_{i=1}^n (y_i - \bar{y}) (x_i - \bar{x}) - \hat{a}^2 \sum_{i=1}^n (x_i - \bar{x})^2
\end{aligned}$$

On montre que cette dernière quantité est nulle en remplaçant $\hat{a}$ par son expression dans la question 2.

Exercice 2 :

Lorsque que l'on fait de la regression linéaire (comme dans le reste des mathématiques d'ailleurs) il faut faire attention aux notations de l'énoncé. Une notation classique de l'équation de régression est en effet

$$\hat{Y} = \hat{a} + \hat{b}X$$

Remarquer que le rôle de $\hat{a}$ et $\hat{b}$ est inverser par rapport à l'exercice 1. Dans ce cas,

$$\hat{b} = \frac{\sum_{i=1}^n (x_i y_i) - n\bar{x}\bar{y}}{\sum_{i=1}^n (x_i^2) - n\bar{x}^2}, \quad \hat{a} = \bar{y} - \hat{b}\bar{x}.$$

Pour aller plus loin, on pourrait intervertir les rôles de X et Y et vous demander d'écrire X en fonction de Y . L'équation de régression devient alors

$$\hat{X} = \hat{a}Y + \hat{b}$$

où les coefficients $\hat{a}$ et $\hat{b}$ sont donnés par

$$\hat{a} = \frac{\sum_{i=1}^n (x_i y_i) - n\bar{x}\bar{y}}{\sum_{i=1}^n (y_i^2) - n\bar{y}^2}, \quad \hat{b} = \bar{x} - \hat{a}\bar{y}.$$

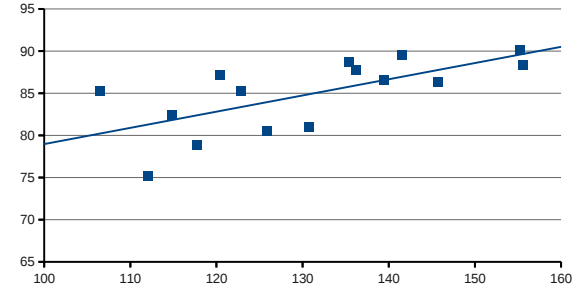


FIG. 1 – Nuage de points associé au tableau de l'exercice 3 et sa courbe de régression linéaire d'équation $f(x) = 0.19x + 59.78$.

Exercice 3 :

1. Voir la figure 1.
2. On note X le poids des feuilles et Y le poids des racines. L'équation de la régression linéaire s'écrit $\hat{Y} = \hat{a}X + \hat{b}$ où $\hat{a}$ et $\hat{b}$ sont données par les formules de l'exercice 1. Les applications numériques donnent $\hat{a} = 0.19$ et $\hat{b} = 59.78$.
3. Le coefficient de corrélation est donné par la formule de l'exercice 1. L'application numérique donne $r = 0.67$. Le coefficient de détermination est alors le carré du coefficient de corrélation, c'est à dire $r^2 = 0.45$. On en déduit que les variations de Y sont moyennement expliquées par celles de X . Noter que les calculs confirment l'impression visuelle.
4. L'équation de la courbe de régression linéaire est $f(x) = 0.19x + 59.78$. Lorsque que le poids des feuilles est de 95 grammes, le poids des racines serait de $f(95) = 78.0$ grammes. De même pour 110 (resp. 130) grammes de feuilles, on auraient 80.9 (resp. 88.6) grammes de racines.

Exercice 4 :

1. L'application numérique donne $r = 0$ (arrondi à 10^{-2} près).
2. Non, on peut seulement en conclure que s'il y a un lien, ce lien n'est pas linéaire.
3. Voir la figure 2.

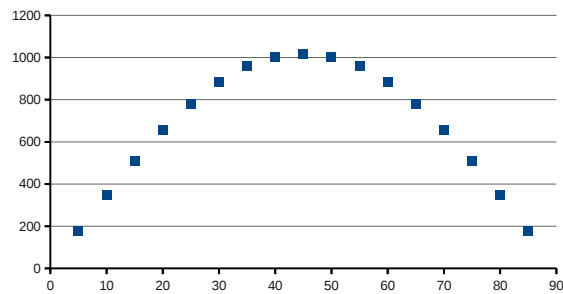


FIG. 2 – Nuage de points associé au tableau de l'exercice 4.

4. Clairement, il y a un lien entre les deux variables. Plus précisément, Y semble pouvoir s'écrire sous la forme $aX^2 + bX + c$ avec des coefficients a , b et c bien choisis.