# Exercices - Série 3
# Régression linéaire simple

## Exercice 1 - Densité européenne

Dans le fichier *Europe.xlsx* disponible sur le site web du cours, vous trouverez la population et la superficie de 27 pays d'Europe.

a) Tracer le nuage de points de la population en fonction de la superficie. En examinant ce graphique, êtes-vous portés à dire que les postulats du modèle de régression linéaire sont respectés ? Quelle est la conséquence de votre réponse ?

b) Pour estimer la densité de population moyenne en Europe, on propose 3 approches :

   i) en calculant la densité de chaque pays, puis en faisant la moyenne de ces 27 densités.
   ii) en calculant la population totale des 27 pays, et en la divisant par la superficie totale des 27 pays.
   iii) en estimant la pente de la droite de régression aux moindres carrés

   Exprimer les 3 quantités ci-dessus en nombre d'habitants par km$^2$, et commenter chacune d'elles. Laquelle des approches vous apparaît la meilleure ?

## Exercice 2 - Drill, baby, drill ! (Comme disait Sarah Palin)

a) Montrer que la somme des produits croisés $S_{XY} = \sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})$ peut aussi s'écrire $S_{XY} = \sum_{i=1}^{n} X_i Y_i - n\overline{X}\,\overline{Y}$.

b) Montrer que la somme des produits croisés $S_{XY} = \sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})$ peut aussi s'écrire $S_{XY} = \sum_{i=1}^{n}(X_i - \overline{X})Y_i$.

c) Retrouver les estimateurs des moindres carrés de la pente et de l'ordonnée à l'origine en annulant les dérivées partielles par rapport à $\beta_0$ et $\beta_1$ de la somme des carrés des erreurs :

$$S = \sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)^2$$

d) On peut déduire des résultats précédents que $\hat{\beta}_1$ est une combinaison linéaire des $Y_i$ (lorsqu'on considère les $X_i$ fixés). Quelle est la principale conséquence de cet état de fait ?

**Exercice 3 - Dans le ventre de sa maman...**

Téléchargez sur le site du cours le jeu de données *gestation_longévité.xlsx*. À l'aide de la commande *Régression linéaire* de l'*Utilitaire d'analyse*, ajustez un modèle de régression linéaire simple sur les variables suivantes : (Demandez de faire calculer les résidus.)

| | | | |
|---|---|---|---|
| Modèle 1 : | Longévité | en fonction de | Gestation |
| Modèle 2 : | Longévité | en fonction de | ln(Gestation) |
| Modèle 3 : | ln(Longévité) | en fonction de | Gestation |
| Modèle 4 : | ln(Longévité) | en fonction de | ln(Gestation) |

a) En étudiant les quatre nuages de points, lequel des modèles proposés est préférable pour mettre en lien ces deux variables ? Pourquoi ?

b) Dans les résultats de son analyse, Excel fournit trois coefficients permettant d'évaluer la qualitéde l'ajustement. Pouvez-vous les identifier, et préciser la formule utilisée pour calculer chacun d'eux ?

c) En vous basant sur le $R^2$, lequel des quatre modèles proposés fournit le meilleur ajustement ?

d) Pour le modèle que vous avez choisi en a), quelle est l'estimation de la variance des observations autour de la droite ?

e) Pour le modèle que vous avez choisi en a), calculez la moyenne et l'écart-type des résidus. Auriez-vous pu trouver ces valeurs sans les faire calculer par Excel à partir de la liste des résidus ?

**Exercice 4 - Jouons avec les Y**

Considérons les données suivantes :

| Variable $X$ | Variable $Y_1$ | Variable $Y_2$ | Variable $Y_3$ |
|---|---|---|---|
| 0 | 1 | 11,0 | 1,5 |
| 10 | 1,5 | 10,4 | 2,6 |
| 19 | 2,6 | 8,3 | 5,1 |
| 27 | 2,7 | 7,9 | 7,9 |
| 29 | 5,0 | 7,2 | 7,2 |
| 32 | 5,1 | 5,1 | 10,4 |
| 34 | 7,2 | 5,0 | 2,7 |
| 43 | 7,9 | 2,7 | 11,0 |
| 50 | 8,3 | 2,6 | 8,3 |
| 51 | 10,4 | 1,5 | 5,0 |
| 55 | 11,0 | 1,0 | 1,0 |

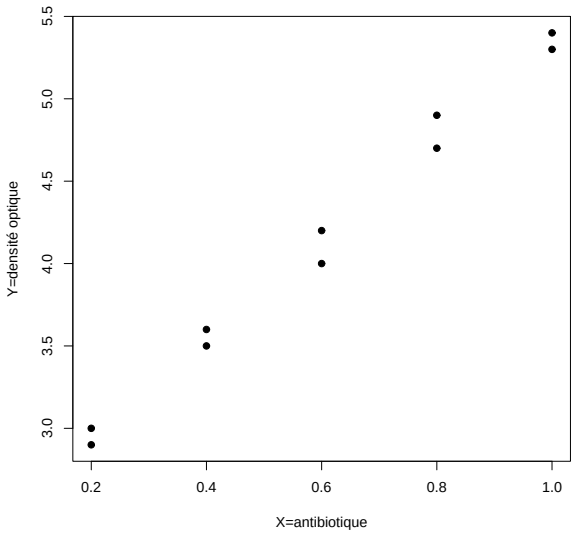
a) Estimer les paramètres du modèle de régression linéaire de $Y_1$ en fonction de $X$ :
   i) par la méthode de Mayer ;
   ii) par la méthode médiane-médiane ;

b) Si on inverse l'ordre des $Y_i$ (et donc qu'on travaille avec la variable $Y_2$), qu'advient-il des paramètres avec les deux méthodes ? Pouvez-vous répondre sans faire de calcul ?

c) Si on change complètement l'ordre des $Y_i$ (et donc qu'on travaille avec la variable $Y_3$), devez-vous réordonner les $Y_i$ avant de calculer les points moyens et médians ?

**Exercice 5 - Un air de déjà vu...**

Un scientifique désire étudier l'influence d'un antibiotique sur une culture bactérienne. Il répartit dans 10 tubes des volumes égaux de culture additionnée d'une quantité $X$ d'antibiotique, et il mesure (après incubation) la densité optique $Y$. Il a noté que $\overline{x} = 0,6$ et que $\overline{y} = 4,15$.
Les résultats sont illustrés par le graphique suivant :




| $X$ = antibiotique | 0.2 | 0.2 | 0.4 | 0.4 | 0.6 |
|---|---|---|---|---|---|
| $Y$ = densité | 2.9 | 3.0 | 3.5 | 3.6 | 4.0 |

| $X$ = antibiotique | 0.6 | 0.8 | 0.8 | 1.0 | 1.0 |
|---|---|---|---|---|---|
| $Y$ = densité | 4.2 | 4.7 | 4.9 | 5.3 | 5.4 |

a) D'après les résultats de l'expérience, lorsqu'il n'y a pas d'antibiotique, la densité optique est estimée à 2,335. Trouver l'équation de la droite de régression. Ajouter la droite de régression à votre graphique.

b) Sachant que la variance échantillonnale des $x_i$ est 0,0889, que celle des $y_i$ est 0,8206, et que la somme des carrés des erreurs du modèle est de 0,0645, donner une estimation de l'erreur-type associée à $\hat{\beta}_0$ et de l'erreur-type associée à $\hat{\beta}_1$.

c) Peut-on dire que la pente de la droite est supérieure à 3 au seuil $\alpha = 5\%$ ?

d) Quelle est la proportion de variation de la densité optique expliquée par la quantité d'antibiotique ?

e) Un onzième tube est additionné d'une quantité d'antibiotique égale à 0,9. Donner un intervalle de valeurs permettant de prédire la densité optique de ce tube qu'on mesurera après incubation. Utiliser $\alpha = 0,05$.

f) Toutes choses étant égales par ailleurs, l'intervalle que vous avez calculé au numéro précédent aurait-il été plus long ou plus court...

   i) si on avait choisi une quantité d'antibiotique égale à 0,7 ?

   ii) si on avait choisi $\alpha = 0,01$ ?

   iii) si on avait utilisé une taille d'échantillon de 20 unités ?

   iv) si on avait construit l'intervalle pour estimer la densité optique moyenne de tous les tubes ayant reçu une quantité d'antibiotique égale à 0,9 ?

g) Considérons un tube dont la quantité d'antibiotique additionnée est située à 1,5 écart-type sous la moyenne.

   i) Combien d'unités d'antibiotique ce tube a-t-il reçu ?

   ii) Quelle est la valeur de densité optique prédite par le modèle pour ce tube ?

   iii) À combien d'écarts-types de la densité optique moyenne cette valeur se situe-t-elle ? Est-ce à 1,5 écart-type sous la densité moyenne ?

   iv) Pouvez-vous déduire de vos calculs au numéro précédent la valeur du coefficient de corrélation ?

h) Un autre analyste propose de calculer la moyenne de tous les $y_i$ ayant une même valeur de $X$ avant d'ajuster un modèle de régression. On utiliserait donc les cinq points suivants :

| $X =$ antibiotique | 0.2 | 0.4 | 0.6 | 0.8 | 1.0 |
|---|---|---|---|---|---|
| $Y =$ densité | 2.95 | 3.55 | 4.1 | 4.8 | 5.35 |

   Pouvez-vous déterminer quel sera l'impact de cette approche sur ...

   i) la moyenne des $x_i$ et des $y_i$ ?
   ii) l'équation de la droite de régression ?
   iii) l'estimation de la variance autour de la droite et la précision des prédictions ?

   L'une des deux analyses est-elle préférable à l'autre ?

### Exercice 6 - Ma cabane au Canada

On veut modéliser le prix moyen des maisons au Canada en fonction du temps et on obtient les données suivantes (malheureusement incomplètes...) :

| Année ($X$) | 1980 | 1981 | ... | 2010 |
|---|---|---|---|---|
| Prix moyen des maisons (\$) ($Y$) | 74 721 | 76 236 | ... | 249 017 |

Une analyse de ces données montre qu'un modèle linéaire pourrait éventuellement s'appliquer pour expliquer le prix des maisons à partir de l'année. On calcule la covariance et la corrélation échantillonnales, et on obtient les quantités suivantes :

$$\begin{aligned} Cov(X, Y) &= 374\,225 \\ r &= 0,77 \end{aligned}$$

a) On décide d'exprimer le prix des maisons en milliers de dollars plutôt qu'en dollars. Que deviennent la covariance et la corrélation ?

b) Conservons les prix initiaux (en dollars). On veut maintenant exprimer le temps en nombre d'années écoulées depuis 1980. Qu'advient-il de la covariance et de la corrélation ?