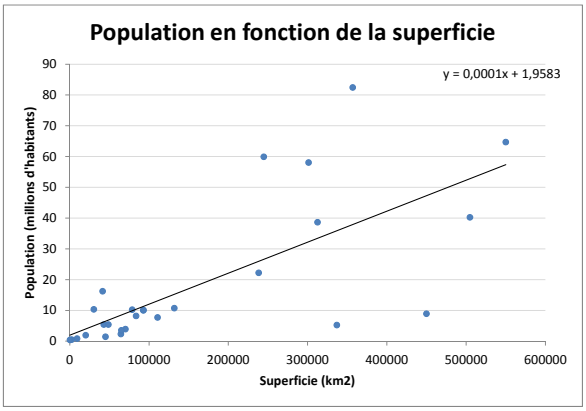
Corrigé - Série 3
Régression linéaire simple

Exercice 1 - Densité européenne

a)



On voit qu'il y a probablement une relation linéaire croissante entre la population et la superficie. Par contre, il est clair que la variance n'est pas constante autour de la droite (les résidus afficheraient un entonnoir ouvert à droite). On peut donc ajuster un modèle linéaire avec n'importe quelle méthode d'estimation (calculer l'équation d'une droite), mais on ne peut pas associer de marge d'erreur aux estimations des moindres carrés comme on le ferait si tous les postulats étaient respectés.

b) Estimation de la densité moyenne de la population en Europe :

i) en calculant la moyenne des 27 densités :

$$\frac{\sum_{i=1}^{27} y_i/x_i}{27} = 166,28 \text{ hab/km}^2$$

Ce calcul donne un poids égal à chaque pays. C'est la moyenne des densités des pays d'Europe, donc c'est la densité moyenne par pays. Les petits pays, ayant souvent une grande densité, ont plus de poids dans ce calcul.

ii) en calculant la population totale des 27 pays, et en la divisant par la superficie totale des 27 pays :

$$\sum_{i=1}^{27} y_i \Big/ \sum_{i=1}^{27} x_i = 112,95 \text{ hab/km}^2$$

Ce calcul donne un poids égal à chaque km^2 de territoire. Les grands pays ont plus de poids dans ce calcul. Cette formule ne tient pas compte des divisions

politiques. Si l'Europe était un pays, ce serait sa densité de population. Bien sûr, cette densité n'est pas homogène.

iii) en estimant la pente de la droite de régression aux moindres carrés :

$$\frac{\sum_{i=1}^{27} x_i y_i - 27\overline{x}\,\overline{y}}{\sum_{i=1}^{27} x_i^2 - 27\overline{x}^2} = 100,74 \text{ hab/km}^2$$

Ce calcul donne une estimation de l'augmentation moyenne de la population lorsque le territoire augmente d'un km^2. Cette estimation ne correspond pas exactement à la valeur en a), car elle est calculée en minimisant l'erreur de prédiction de la population à partir d'une superficie connue (les distances verticales par rapport à la droite).

Si la droite passait par 0 exactement, ce serait une façon d'envisager la densité "moyenne" (et on n'en est pas loin, puisque $\hat{\beta}_0 = 1,96$). À titre informatif, on peut forcer la droite de régression à passer par 0 (en minimisant la somme du carré des erreurs du modèle $Y_i = \beta_1 x_i + \varepsilon_i$), on obtient alors l'estimation suivante pour la pente :

$$\frac{\sum_{i=1}^{27} x_i y_i}{\sum_{i=1}^{27} x_i^2} = 106,84$$

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

a)

$$\begin{aligned}
S_{XY} &= \sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}) \\
&= \sum_{i=1}^{n}(X_iY_i - X_i\overline{Y} - Y_i\overline{X} + \overline{XY}) \\
&= \sum_{i=1}^{n} X_iY_i - \overline{Y}\sum_{i=1}^{n} X_i - \overline{X}\sum_{i=1}^{n} Y_i + n\overline{XY} \\
&= \sum_{i=1}^{n} X_iY_i - \overline{Y}(n\overline{X}) - \overline{X}(n\overline{Y}) + n\overline{XY} \\
&= \sum_{i=1}^{n} X_iY_i - n\overline{X}\,\overline{Y}
\end{aligned}$$

b)

$$\begin{aligned} S_{XY} &= \sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}) \\ &= \sum_{i=1}^{n}(X_i - \overline{X})Y_i - \sum_{i=1}^{n}(X_i - \overline{X})\overline{Y} \\ &= \sum_{i=1}^{n}(X_i - \overline{X})Y_i - \overline{Y}\sum_{i=1}^{n}(X_i - \overline{X}) \\ &= \sum_{i=1}^{n}(X_i - \overline{X})Y_i - \overline{Y}(0) \\ &= \sum_{i=1}^{n}(X_i - \overline{X})Y_i \end{aligned}$$

c) $\dfrac{\partial S}{\partial \beta_0} = 0$ si $\sum_{i=1}^{n} Y_i = n\,\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} X_i$

On isole $\hat{\beta}_0$ et on obtient $\boxed{\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}}$.

$\dfrac{\partial S}{\partial \beta_1} = 0$ si $\sum_{i=1}^{n} X_i Y_i = \hat{\beta}_0 \sum_{i=1}^{n} X_i + \hat{\beta}_1 \sum_{i=1}^{n} X_i^2$

En remplaçant $\hat{\beta}_0$ par $\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$, on obtient :

$$\sum_{i=1}^{n} X_i Y_i - \overline{Y}\sum_{i=1}^{n} X_i = \hat{\beta}_1(\sum_{i=1}^{n} X_i^2 - \overline{X}\sum_{i=1}^{n} X_i)$$

$$\sum_{i=1}^{n} X_i Y_i - n\overline{X}\,\overline{Y} = \hat{\beta}_1(\sum_{i=1}^{n} X_i^2 - n\overline{X}^2)$$

On isole $\hat{\beta}_1$ et on obtient $\boxed{\hat{\beta}_1 = \dfrac{S_{XY}}{S_{XX}}}$

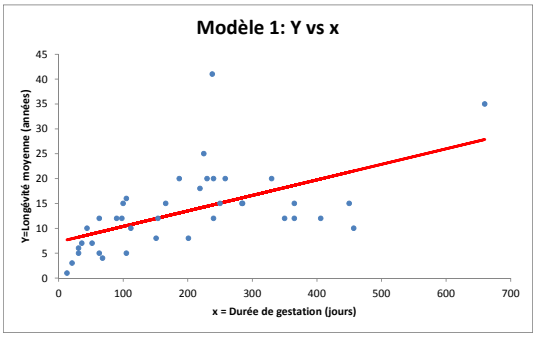
d) En effet, $\hat{\beta}_1 = \dfrac{S_{XY}}{S_{XX}} = \dfrac{\sum_{i=1}^{n}(X_i - \overline{X})Y_i}{\sum_{i=1}^{n}(X_i - \overline{X})^2}$

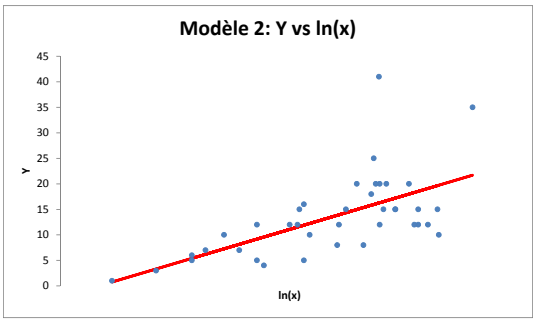
La principale conséquence de cet état de fait est que $\hat{\beta}_1$ suit une loi normale lorsqu'on suppose que les Y_i suivent une loi normale (autour de la droite).

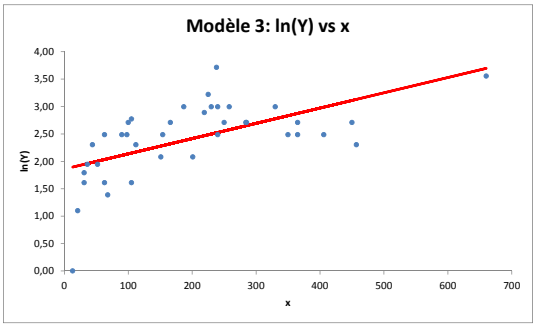
Exercice 3 - Dans le ventre de sa maman...

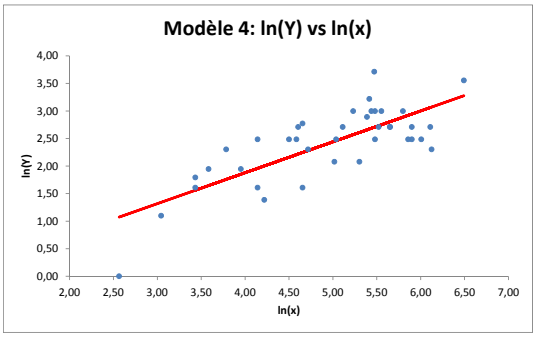
Modèle 1 :	Longévité	en fonction de	Gestation
Modèle 2 :	Longévité	en fonction de	ln(Gestation)
Modèle 3 :	ln(Longévité)	en fonction de	Gestation
Modèle 4 :	ln(Longévité)	en fonction de	ln(Gestation)

a) Selon les quatre graphiques de dispersion, le modèle 4 est clairement celui qui présente la relation la plus linéaire, avec une variance à peu près constante pour toutes les valeurs de x.









b)

Appellation dans Excel	Symbole	Formule
`Coeff. de détermination multiple`	r	coeff. de corrélation échantillonnal $\frac{Cov(X,Y)}{S_X \cdot S_Y} = \frac{S_{XY}}{\sqrt{S_{XX} \cdot S_{YY}}}$
`Coeff. de détermination R^ 2`	R^2	$1 - \frac{SSE}{SST} = \frac{SSR}{SST} = r^2$
`Coeff. de détermination R^ 2`	R^2_{ajuste}	$1 - \frac{SSE/(n-2)}{SST/(n-1)} = 1 - \frac{MSE}{S_y^2}$

c)

Modèle 1 :	Y	en fonction de	X	$R^2 = 0.3275$
Modèle 2 :	Y	en fonction de	ln(X)	$R^2 = 0.3925$
Modèle 3 :	ln(Y)	en fonction de	X	$R^2 = 0.3535$
Modèle 4 :	ln(Y)	en fonction de	ln(X)	$R^2 = 0.5883$

Le modèle 4 est encore privilégié, car c'est celui pour lequel la proportion de variabilité expliquée par le modèle est la plus grande.

d) $\sigma^2 = MSE = 0.2000$

e) moyenne des résidus $= -3.47 \times 10^{-16} \approx 0$ et écart-type des résidus $= 0.4413$.

On aurait pu trouver ces valeurs sans utiliser la liste des résidus, car la moyenne des écarts est toujours 0, et la variance échantillonnale des résidus correspond à une petite transformation du MSE, soit

$$s_\varepsilon^2 = \frac{\sum_{i=1}^{n}(\hat{\varepsilon}_i - \overline{\varepsilon})^2}{(n-1)} = \frac{\sum_{i=1}^{n}([y_i - \hat{y}_i] - 0)^2}{(n-1)} = \frac{(n-2)MSE}{(n-1)}$$

Exercice 4 - Jouons avec les Y

a) i) Méthode de Mayer :
Deux points moyens : $P_1 = (19,5;\ 3,0)$ et $P_2 = (44,17;\ 8,3)$
Équation de la droite : $\hat{Y}_1 = 0,2162\,x - 1,2329$

ii) Méthode médiane-médiane :
Trois points médians : $P_1 = (14,5;\ 2,1)$, $P_2 = (32;\ 5,1)$ et $P_3 = (50,5;\ 9,4)$
Moyenne des points médians : $(32,33;\ 5,50)$
Équation de la droite : $\hat{Y}_1 = 0,2028\,x - 1,0565$

b) La pente changera de signe, mais aura la même valeur absolue. Pour l'ordonnée à l'origine, les calculs sont nécessaires :

i) Équation de la droite de Mayer : $\hat{Y}_2 = -0,2162\,x + 12,5329$

ii) Équation de la droite médiane-médiane : $\hat{Y}_2 = -0,2028\,x + 12,0565$

c) Non, les valeurs de Y sont liées aux valeurs de X. On ne peut pas séparer les valeurs d'un même individu. On ordonne selon X et les Y suivent.

Exercice 5 - Un air de déjà vu...

a) On sait que la droite de régression passe par $(\overline{x}, \overline{y}) = (0,6, 4,15)$. On donne un autre point dans la question, soit $(0, \hat{\beta}_0) = (0,\ 2,335)$. On peut donc évaluer la pente de la droite :

$$\hat{\beta}_1 = \frac{\Delta y}{\Delta x} = \frac{4,15 - 2,335}{0,6 - 0} = 3,025$$

L'équation de la droite de régression : $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x = 2,335 + 3,025x$

b) $S_{XX} = (n-1)s_X^2 = 9 \times 0,0889 = 0,8001$

$S_{YY} = (n-1)s_Y^2 = 9 \times 0,8206 = 7,388$

$MSE = SSE/(n-2) = 0,0645/8 = 0,00806$

Erreur-type$(\hat{\beta}_0)$: $\sqrt{MSE\left(\frac{1}{n} + \frac{\overline{x}^2}{S_{XX}}\right)} = \sqrt{0,00806\left(\frac{1}{10} + \frac{0,6^2}{0,8001}\right)} = 0,0666$

Erreur-type$(\hat{\beta}_1)$: $\sqrt{\frac{MSE}{S_{XX}}} = \sqrt{\frac{0,00806}{0,8001}} = 0,1004$

c) Il y a deux façons de répondre à cette question. Il s'agit d'un test unilatéral sur β_1 :

$$\begin{aligned} H_0 :&\quad \beta_1 = 3 \\ H_1 :&\quad \beta_1 > 3 \end{aligned}$$

1) Test d'hypothèse sur la pente :

On construit une statistique de Student en se basant sur le fait que $\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_{XX}}\right)$.

Si H_0 est vraie, alors $T_0 = \frac{\hat{\beta}_1 - 3}{err. - type(\hat{\beta}_1)} = \frac{\hat{\beta}_1 - 3}{\sqrt{MSE/S_{XX}}} \sim t_{n-2}$.

Puisque $t_{obs} = \frac{3.025 - 3}{0,1004} = 0.249$ n'est pas supérieure à la valeur critique $t_{\alpha;n-2} = t_{0.05;8} = 1,86$, on ne rejette pas H_0.

2) Intervalle de confiance sur la pente :

Un intervalle de confiance de niveau $1-\alpha$ est équivalent à un test *bilatéral* de seuil α sur un paramètre, car il a deux bornes. La zone de rejet du test unilatéral $(H_1 : \beta_1 > 3)$

exclut les 5% des valeurs les plus improbables de la distribution de $\hat{\beta}_1$ sous H_0 à l'extrémité *droite* du spectre. L'intervalle de confiance correspondant devra "exclure" 5% des valeurs à chaque extrémité du spectre. Il aura donc un niveau de 90%.

$$\hat{\beta}_1 \pm t_{8;0,05} \times err. - type(\hat{\beta}_1) = 3,025 \pm 1,86 \times 0,1004 = [2,838,\ 3,212]$$

On rejetterait H_0 si toutes les valeurs de l'intervalle de confiance étaient supérieures à 3. C'est donc la borne inférieure qui détermine notre décision. On ne peut donc pas conclure que la pente de la droite est supérieure à 3 au seuil $\alpha = 5\%$.

d) $R^2 = \dfrac{SSR}{SST} = 1 - \dfrac{SSE}{SST} = 1 - \dfrac{SSE}{S_{YY}} = 1 - \dfrac{0,0645}{7,388} = 0,9913.$

e) Intervalle de prédiction pour une observation future :

$$\begin{aligned} \hat{y}_0 &\pm t_{8;0,025}\sqrt{MSE\left(1 + \tfrac{1}{n} + \tfrac{(x_0-\overline{x})^2}{S_{XX}}\right)} \\ 2,335 + 3,025(0,9) &\pm 2.306\sqrt{0.00806\left(1 + \tfrac{1}{10} + \tfrac{(0,9-0,6)^2}{0,8001}\right)} \\ 5,058 &\pm 0,228 \\ [4,830 &\ ,\ 5,286] \end{aligned}$$

f) i) Si on avait choisi une quantité d'antibiotique égale à 0,7, l'intervalle aurait été plus court car x_0 aurait été plus proche de la moyenne.

ii) Si on avait choisi $\alpha = 0.01$, l'intervalle aurait été plus long, car le quantile $t_{8;0,005}$ aurait été plus grand.

iii) Si on avait utilisé une taille d'échantillon de 20 unités, l'intervalle aurait été plus court, car n et S_{XX} seraient plus grands et $t_{n-2;0,025}$ serait plus petit.

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, l'intervalle aurait été plus court, car on aurait choisi la formule de l'intervalle de confiance pour $E(Y|x_0)$, qui tient seulement compte de l'erreur d'estimation du point sur la droite.

g) i) Ce tube a reçu $\overline{x} - 1,5s_X = 0,6 - 1,5(0,298) = 0,153$ unité d'antibiotique.

ii) densité optique prédite = 2,797

iii) Cette valeur se situe à 1,493 écart-type de la densité optique moyenne

iv) $1,5 \times r = 1,493 \quad \rightarrow \quad r = 0,995$. On peut vérifier qu'il s'agit de la racine carrée positive de R^2.

h)

i) La moyenne des x_i et des y_i restera exactement la même dans les deux cas.

ii) L'équation de la droite de régression restera inchangée.

On peut voir facilement que S_{xx} sera deux fois plus petit, car chaque écart est présent une seule fois dans la somme au lieu de deux.

S_{xy} sera aussi deux fois plus petit. Pour s'en convaincre, prenons les deux premiers termes de la somme avant réduction des données :

$$(0.2 - \overline{x})(2.9 - \overline{y}) + (0.2 - \overline{x})(3.0 - \overline{y}) = (0.2 - \overline{x})(2.9 + 3.0 - 2\overline{y})$$

Ils sont maintenant remplacés par $(0.2 - \overline{x})(2.95 - \overline{y})$ dans la somme après réduction des données. Idem pour les huit autres termes de S_{xy}.

iii) L'estimation de la variance autour de la droite sera considérablement réduite et par conséquent la marge d'erreur sur les prédictions sera faussement diminuée. Une bonne partie de la variabilité naturelle dans les observations est occultée par la mise en commun des Y ayant la même valeur de X.

Exercice 6 - Ma cabane au Canada

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

a) On décide d'exprimer le prix des maisons en milliers de dollars plutôt qu'en dollars. Posons $W = Y/1\,000$.

$$Cov(X, W) = Cov\left(X, \frac{Y}{1\,000}\right) = \frac{1}{1\,000} Cov(X, Y) = 374,\ 225$$

$$r(X, W) = \frac{Cov(X, W)}{\sqrt{Var(X)\ Var(W)}} = \frac{Cov(X, Y/1\,000)}{\sqrt{Var(X)\ Var(Y/1\,000)}} = \frac{1/1\,000\ Cov(X, Y)}{\sqrt{Var(X)\ (1/1\,000)^2 Var(Y)}}$$

$$= \frac{Cov(X, Y)}{\sqrt{Var(X)\ Var(Y)}} = r(X, Y) = 0,77$$

b) On veut maintenant exprimer le temps en nombre d'années écoulées depuis 1980. Posons $T = X - 1980$.

$$Cov(T, Y) = Cov\left(X - 1980, Y\right) = Cov(X, Y) = 374\,225$$

$$r(T, Y) = \frac{Cov(X - 1980, Y)}{\sqrt{Var(X - 1980)\ Var(Y)}} = \frac{Cov(X, Y)}{\sqrt{Var(X)\ Var(Y)}} = r(X, Y) = 0,77$$