IDUP / avril 21

Master 1 : Analyse statistique (JF Léger)

Mesure du lien entre deux variables quantitatives : l'ajustement linéaire

Support de cours

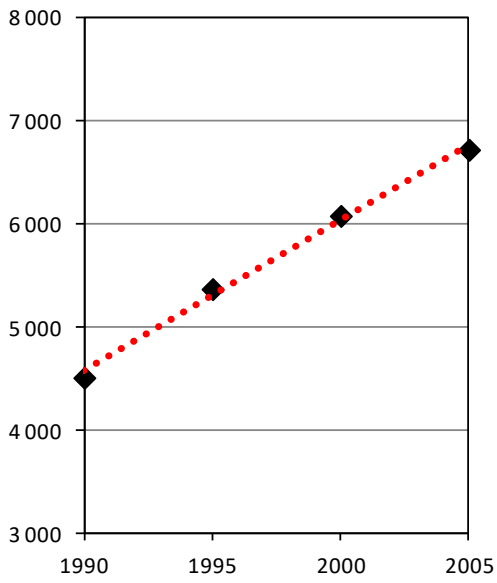
Objectif de l'ajustement linéaire

On réalise un ajustement linéaire pour mesurer la force d'un lien statistique entre deux variables quantitatives. D'une manière générale, la réalisation d'un graphique permet de mettre en évidence l'existence d'un tel lien. L'ajustement linéaire permet de modéliser ce lien au moyen d'une fonction mathématique qui lie la variable explicative (notée *x*) et la variable expliquée ou dépendante, c'est-à-dire celle qui dépend de l'autre (notée *y*) : y = f(x)

Selon la forme du nuage de point, la fonction qui lie les variables *x* et *y* varie. Le cas de référence est celui où les points sont situés à proximité d'une droite « imaginaire » (figure 1). Dans ce cas, la fonction reliant *x* à *y* est une droite et la relation entre *x* et *y* peut être formalisée de la manière suivante : **y** = a**x**+ b où :

- *a* correspond à la pente de la droite dite de régression linéaire. Il s'agit de la variation absolue de *y* par unité de *x* ;
- *b* est la valeur que prend *y* quand *x* vaut 0. C'est la raison pour laquelle on appelle cette valeur constante l'ordonnée (c'est-à-dire la valeur que prend *y*) à l'origine (qui correspond à *x* = 0).

Figure 1 : Evolution de l'effectif de la population israélienne (en milliers) entre 1990 et 2005 (en pointillés la droite d'ajustement linéaire)



Mais le nuage de points peut avoir une forme différente (figures 2 et 3). Le lien mathématique entre les variables x et y s'exprime alors différemment. Par exemple :

- le cumul des naissances pour 100 femmes de la génération 1947 (descendance atteinte selon l'âge ; figure 2) peut être modélisé par une fonction logarithmique (**y** = aLn(**x**) + b) ;
- l'évolution des risques de mortalité selon l'âge (figure 3) peut être ajustée par un modèle de type exponentiel (**y** = exp (a**x** + b)).

Figure 2 : Descendance atteinte des femmes de la génération 1947 selon l'âge (pour 100 femmes) et ajustement

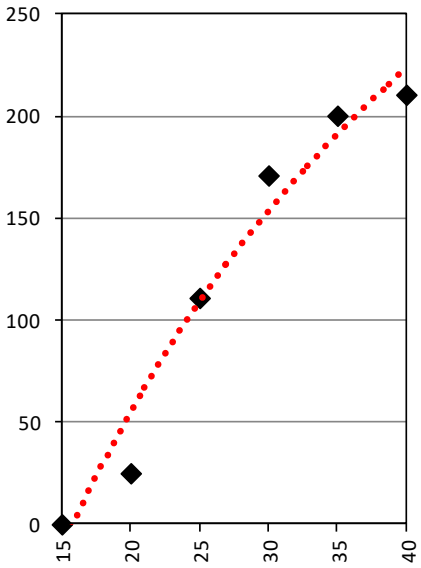
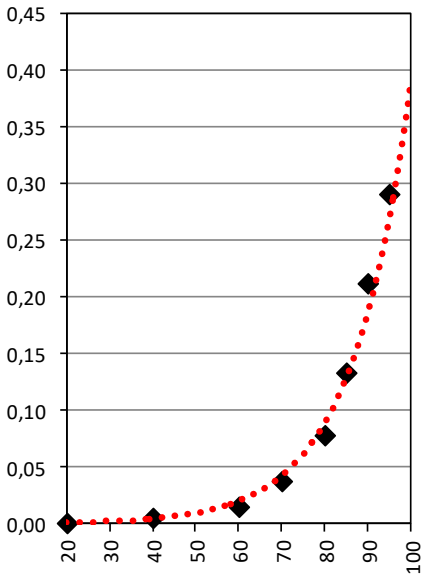


Figure 3 : Risques de mortalité selon l'âge à partir de 20 ans (hommes + femmes) en France en 2003 et ajustement



Tous ces modèles (linéaire, logarithmique, exponentiel) présentent la caractéristique de formaliser un lien monotone (croissant ou décroissant) entre deux variables sur un intervalle (d'âge, de temps, etc.) donné : quand l'une diminue (ou augmente), l'autre ne cesse d'augmenter ou de diminuer (et vice versa).

Des relations plus complexes peuvent exister entre variables quantitatives. On peut modéliser ces relations avec des fonctions polynomiales. Plus simplement, on peut réaliser plusieurs ajustements en scindant la série statistique en plusieurs séquences.

On peut réaliser automatiquement de nombreux types d'ajustements avec Excel©, qui propose, en plus des modèles linéaire, logarithmique et exponentiel, des ajustements au moyen de fonctions polynomiales de degré n. A chaque fois, il est possible de mesurer la force du lien à l'aide du coefficient de détermination et du coefficient de corrélation linéaire. Excel permet également de réaliser à partir de ces modèles des extrapolations ou des rétropolations.

Principe de l'ajustement linéaire

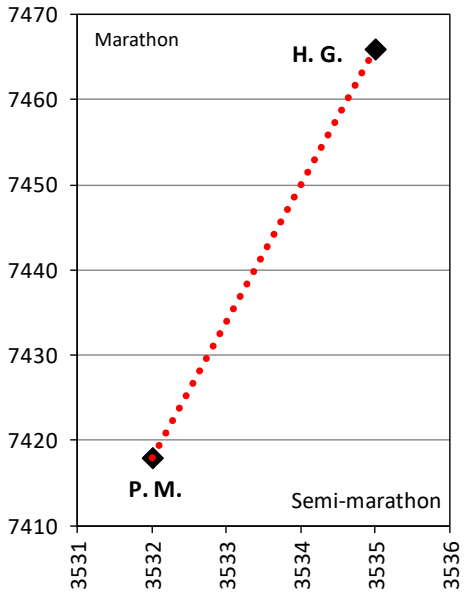
On va prendre comme exemple la relation entre le temps mis par certains champions de course à pied pour parcourir un semi-marathon et un marathon. On va commencer par considérer cette relation à partir des performances de deux coureurs, le recordman du monde du marathon en 2014, Patrick Makam et son prédécesseur, Haile Gebreselassie. Les données du tableau 1 montrent que le temps mis pour parcourir le marathon correspond à peu près pour chacun de ces coureurs à deux fois le temps mis pour courir un semi-marathon plus quelques minutes (environ 5-6 minutes).

Tableau 1 : Performances de Patrick Makam et Haile Gebreselassie au semi-marathon et au marathon

Athlète	Performance en h, mn, sec		Conversion des performances en secondes	
	Semi	Marathon	Semi	Marathon
Patrick Makam	58 min 52 s	2 h 03 min 38 s	3 532 s	7 418 s
Hailé Gebreselassié	58 min 55 s	2 h 04 min 26 s	3 535 s	7 466 s
Différence HG vs. PM			+ 3 s	+ 48 s

On peut préciser cette relation (figure 4) par le calcul. Par deux points passe une et une seule droite. Nous allons chercher les paramètres de cette droite (la pente et l'ordonnée à l'origine). Il s'agit ici de résoudre un système de deux équations à deux inconnues.

Figure 4 : Performances de Patrick Makam et Haile Gebreselassie au semi-marathon et au marathon (temps en seconde)



$$y = a \times x + b$$

$$\begin{cases} PM : 7\,418 = a \times 3\,532 + b \\ HG : 7\,466 = a \times 3\,535 + b \end{cases}$$

$$7\,418 - 7\,466 = a \times 3\,532 + b - \left(a \times 3\,535 + b\right)$$

$$-48 = a \times \left(3\,532 - 3\,535\right)$$

$$a = \frac{-48}{-3} = 16$$

$$b = 7\,418 - 16 \times 3\,532 = -49\,094$$

Pour passer du temps mis pour parcourir un semi-marathon à celui mis pour couvrir un marathon, il faut donc multiplier la performance au semi-marathon (en secondes) par 16 et retrancher 49 094 secondes : $y = 16 \times x - 49\,094$

Sens de la pente (a)

Ces deux coureurs ont réalisé des performances très proches sur le semi-marathon (3 secondes en faveur de Patrick Makam), et plus « éloignées » sur le marathon (48 secondes d'écart, toujours en faveur de Patrick Makam). Une seconde d'écart entre les deux coureurs sur le semi marathon se traduit donc par un écart de 16 secondes sur le marathon. La pente (ici + 16) correspond donc à la variation dans l'unité de mesure utilisée (ici les secondes) de la variable expliquée (le temps sur marathon) quand la valeur de la variable explicative (le temps sur semi-marathon) augmente d'une unité (ici une seconde). Cette relation correspond bien à une suite arithmétique de raison 16, puisque par unité de temps supplémentaire au semi-marathon, on gagne un nombre constant d'unité de temps (ici 16) sur le marathon.

Sens de l'ordonnée à l'origine (b)

On peut écrire la relation entre les temps sur semi-marathon et marathon différemment. On prend pour cela les temps réalisés par Patrick Makam comme référence. Le temps sur marathon de Hailé Gebresselassié est égal au temps sur cette distance de Patrick Makam plus 16 fois la différence entre leurs temps respectifs sur semi-marathon :

$$y_{HG} = y_{PM} + 16 \times (x_{HG} - x_{PM})$$
$$y_{HG} - y_{PM} = 16 \times (x_{HG} - x_{PM})$$
$$\Delta y = 16 \times \Delta x$$

Cette relation indique bien que l'écart (en secondes) sur marathon entre les deux coureurs est égal à 16 fois l'écart entre ces deux mêmes coureurs sur le semi-marathon. Or, ce que l'on modélise est le temps sur marathon de Hailé Gébressélassié en fonction du temps sur semi-marathon (*y* en fonction de *x* et non Δy en fonction de Δx).

$$y_{HG} = 16 \times x_{HG} + (y_{PM} - 16 \times x_{PM})$$
$$y_{HG} = 16 \times x_{HG} + b$$

b est donc une valeur constante qui permet de mettre en évidence la relation entre les variables *x* (le temps sur semi-marathon) et *y* (le temps sur marathon) à partir du calcul de la relation entre les écarts de temps (Δx et Δy) sur ces deux distances entre les deux coureurs. On parle d'ordonnée à l'origine car c'est la valeur que prendrait *y* si *x* était égal à 0. Avec cet exemple, on se rend compte que cette grandeur n'a pas de sens véritablement concret : si le temps sur semi marathon était de 0 seconde (impossible, même dopé), le temps sur marathon serait négatif ! La relation n'a donc de sens que sur un intervalle donné (ici un intervalle de temps), ce qui est le cas de tous les ajustements linéaires.

Remarque : on a écrit précédemment avant de formuler la relation entre les temps mis pour parcourir le semi-marathon et le marathon que l'examen des temps montrait que la performance sur marathon correspondait à deux fois le temps mis pour couvrir un semi-marathon plus quelques minutes (environ 5-6 minutes). On peut retrouver cette relation à partir de $y_{HG} = 16 \times x_{HG} - 49\ 094$

$$\frac{49\ 094}{x_{HG}} = \frac{49\ 094}{3\ 535} = 13{,}9$$
$$y_{HG} = 16 \times x_{HG} - 13{,}9 \times x_{HG} = 2{,}1 \times x_{HG}$$

Le temps sur marathon est donc égal à 2,1 fois le temps sur semi-marathon : à 2 fois le temps sur semi-marathon plus 10 % du temps du semi-marathon. Ces 10 % correspondent à 353,5 secondes, soit un peu moins de 6 minutes.

Mesure de l'ajustement linéaire quand le nombre de couples de points est supérieur à 2

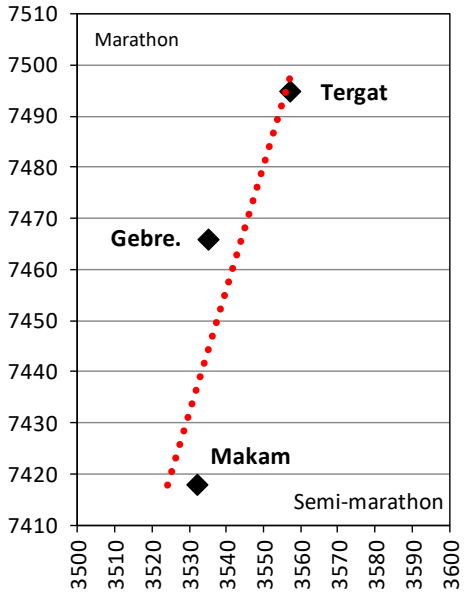
On va ajouter les performances d'un troisième coureur, Paul Tergat, qui a détenu avant les deux athlètes mentionnés ci-dessus le record du monde du marathon.

Tableau 2 : Performances de Patrick Makam, Haile Gebreselassie et Paul Tergat au semi-marathon et au marathon

Athlète	Performance en h, mn, sec		Conversion des perfs. en sec.	
	Semi	Marathon	Semi	Marathon
Patrick Makam	58 min 52 s	2 h 03 min 38 s	3 532 s	7 418 s
Hailé Gebreselassié	58 min 55 s	2 h 04 min 26 s	3 535 s	7 466 s
Paul Tergat	59 min 17 s	2h 04 min 55 s	3 557 s	7 495 s

Les points correspondants à ces trois couples de performances ne sont pas alignés (figure 5). Cela signifie que la relation liant les performances de Patrick Makam et Hailé Gebreselassié ne se vérifie pas pour Paul Tergat. On va tenter de proposer une relation entre ces trois couples de performance la moins mauvaise possible, celle qui minimise en moyenne l'erreur qui sera commise entre la valeur estimée par le modèle (ici le temps théorique sur le marathon) à partir du temps réel sur le semi-marathon et la valeur réelle (le temps réel sur le marathon).

Figure 5 : Performances de Patrick Makam et Haile Gebreselassie au semi-marathon et au marathon (temps en seconde) et ajustement



Il faut déterminer *a* et *b* de telle manière que la droite d'ajustement linéaire s'écarte le moins possible, en moyenne, des points. Un détour par les mathématiques s'impose. Il faut donc que l'écart entre chaque point estimé par le modèle et le point réel soit le plus faible possible. Comme on estime *y* (le temps sur marathon) à partir de *x* réel (le temps sur semi-marathon), il faut donc minimiser l'écart moyen entre la valeur estimée de *y* et la valeur réelle de *y*.

Par définition, la somme des écarts entre valeurs estimées et valeurs réelles est égale à 0. Il faut donc trouver *a* et *b* tels que la somme des carrés des écarts entre les valeurs réelles et estimées (*E*) soit la plus faible possible. Les mathématiciens ont montré que dans le cas d'une fonction de deux variables (ici *a* et *b*, les deux paramètres que l'on cherche à estimer), l'extremum (minimum ou maximum) est déterminé à condition que les dérivés partielles premières de la fonction soient nulles.

Comprendre la démarche mais ne pas la retenir

$$E = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - (a \times x_i + b))^2 = \sum_{i=1}^{n} (y_i - a \times x_i - b)^2$$

$$\begin{cases} \frac{dE}{da} = 0 \\ \frac{dE}{db} = 0 \end{cases}$$

$$E \text{ du type } u^2 \text{ donc } dE = 2u'u$$

$$\begin{cases} \frac{dE}{da} = \sum_{i=1}^{n} 2 \times (-x_i) \times (y_i - a \times x_i - b) = 2\sum_{i=1}^{n} \left(-x_i y_i + a \times x_i^{\,2} + b x_i\right) = -2\sum_{i=1}^{n} \left(x_i y_i - a \times x_i^{\,2} - b x_i\right) = 0 \\ \frac{dE}{db} = \sum_{i=1}^{n} 2 \times (-1) \times (y_i - a \times x_i - b) = -2\sum_{i=1}^{n} (y_i - a \times x_i - b) = 0 \end{cases}$$

Comme :

$$\sum_{i=1}^{n} x_i = n\bar{x} \text{ et } \sum_{i=1}^{n} y_i = n\bar{y}$$

$$\begin{cases} \frac{dE}{da} = 0 \text{ si} \sum_{i=1}^{n} \left(x_i y_i - a \times x_i^{\,2} - b x_i\right) = 0 \\ \frac{dE}{db} = 0 \text{ si} \sum_{i=1}^{n} (y_i - a \times x_i - b) = \sum_{i=1}^{n} y_i - a\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} b = n\bar{y} - an\bar{x} - nb = 0 \end{cases}$$

$$\begin{cases} \frac{dE}{da} = 0 \text{ si} \sum_{i=1}^{n} \left(x_i y_i - a \times x_i^{\,2} - b x_i\right) = 0 \\ \frac{dE}{db} = 0 \text{ si } n(\bar{y} - a\bar{x} - b) = 0 \text{ donc si } b = \bar{y} - a\bar{x} \end{cases}$$

On utilise cette expression de b dans la dérivée de E par rapport à a :

$$\frac{dE}{da} = \sum_{i=1}^{n} \left(x_i y_i - a \times x_i^{\,2} - b x_i\right) = \sum_{i=1}^{n} \left(x_i y_i - a \times x_i^{\,2} - (\bar{y} - a\bar{x}) x_i\right)$$

$$\frac{dE}{da} = \sum_{i=1}^{n} \left(x_i y_i - a \times x_i^{\,2} - \bar{y} x_i + a\bar{x} x_i\right) = \sum_{i=1}^{n} x_i y_i - a\sum_{i=1}^{n} x_i^{\,2} - \bar{y}\sum_{i=1}^{n} x_i + a\bar{x}\sum_{i=1}^{n} x_i$$

$$\frac{dE}{da} = \sum_{i=1}^{n} x_i y_i - a\sum_{i=1}^{n} x_i^{\,2} - \bar{y}(n\bar{x}) + a\bar{x}(n\bar{x}) = \sum_{i=1}^{n} x_i y_i - a\sum_{i=1}^{n} x_i^{\,2} - n\bar{y}\bar{x} + na\bar{x}^{2}$$

$$\frac{dE}{da} = 0 \text{ si } \sum_{i=1}^{n} x_i y_i - a\sum_{i=1}^{n} x_i^{\,2} - n\bar{y}\bar{x} + na\bar{x}^{2} = 0 \text{ donc si } a\left(\sum_{i=1}^{n} x_i^{\,2} - n\bar{x}^{2}\right) = \sum_{i=1}^{n} x_i y_i - n\bar{y}\bar{x}$$

$$\frac{dE}{da} = 0 \text{ si } a = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{y}\bar{x}}{\sum_{i=1}^{n} x_i^{\,2} - n\bar{x}^{2}} = \frac{n\,\mathrm{cov}(x,y)}{n\,\mathrm{var}(x)} = \frac{\mathrm{cov}(x,y)}{\mathrm{var}(x)}$$

La covariance (cov(x,y)) est une mesure de l'association entre *x* et *y*. La covariance est :

- positive lorsque *x* et *y* évoluent dans le même sens (quand y augmente avec *x* ou lorsque *y* diminue en même temps que *x*). Exemple : les risques de mortalité augmente avec l'âge à partir de 20 ans (figure 6) ;
- négative lorsque les variations de *x* et *y* sont inverses (l'une augmente quand l'autre diminue, et vice versa). Exemple : l'espérance de vie diminue à mesure que l'âge augmente (figure 7) ;
- nulle (ou quasi-nulle) quand la variation de l'une des variables est sans effet sur la seconde (une variable croît ou diminue tandis que la seconde est constante ou quasi-constante). Exemple : Evolution du nombre de naissances à Paris au cours des années 2003-2006 (figure 8).

Fig. 6 : Risque de mortalité selon l'âge

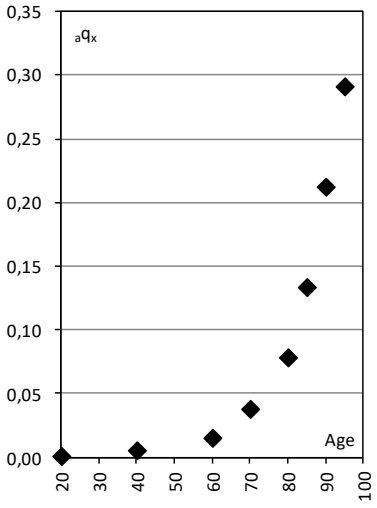


Fig 7 : Espérance de vie selon l'âge

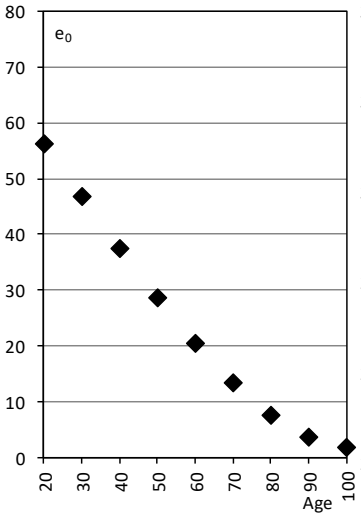
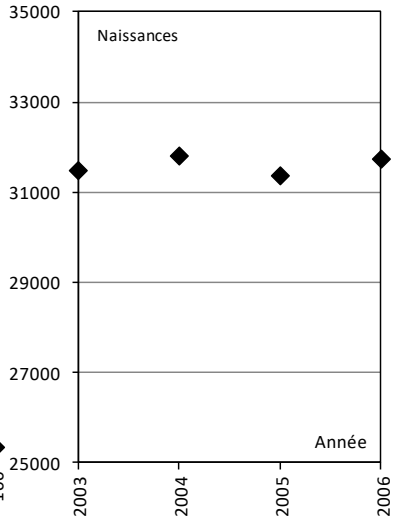


Fig. 8 : Naissances à Paris (2003-2006)



La pente est le rapport entre la covariance et la variance de *x* (var(x)), qui est la moyenne des carrés des écarts à la moyenne. On effectue ce rapport car on détermine la droite de régression des *y* par rapport à *x*.

On peut formaliser la pente de différentes manières :

Connaître les formules

$$a = \frac{\mathrm{cov}(x,y)}{\mathrm{var}(x)} = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\frac{1}{n}\sum_{i=1}^{n}\left[\left(x_i - \bar{x}\right)^2\right]} = \frac{\left[\frac{1}{n}\sum_{i=1}^{n} x_i y_i\right] - \bar{x}\bar{y}}{\left[\frac{1}{n}\sum_{i=1}^{n} x_i^{\,2}\right] - \bar{x}^2} = \frac{\overline{x_i y_i} - \bar{x}\bar{y}}{\overline{x_i^{\,2}} - \bar{x}^2}$$

Et b s'écrit :

$$b = \bar{y} - a\bar{x}$$

Retour à l'exemple des marathoniens. Calcul de a et b

Tableau 3 : Organisation et détails des calculs nécessaires pour déterminer *a* et *b*

Athlète	Semi (x)	Marathon (y)	x - x moy	y - y moy	(x - x moy)*(y - y moy)	(x - x moy)^2
Patrick Makam	3 532	7 418	-9,3	-41,7	387,81	86,49
Haile Gebreselassie	3 535	7 466	-6,3	6,3	-39,69	39,69
Paul Tergat	3 557	7 495	15,7	35,3	554,21	246,49
Somme	10 624	22 379			902,33	372,67
Moyenne	3 541,3	7 459,7			300,777	124,223

$$a = \frac{\text{cov}(x,y)}{\text{var}(x)} = \frac{\frac{1}{3}\sum_{i=1}^{3}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\frac{1}{3}\sum_{i=1}^{3}\left[\left(x_i - \bar{x}\right)^2\right]} = \frac{\frac{1}{3}\left[\left(-9{,}3 \times -41{,}7\right) + \left(-6{,}3 \times 6{,}3\right) + \left(15{,}7 \times 35{,}3\right)\right]}{\frac{1}{3}\left(86{,}49 + 39{,}69 + 246{,}49\right)}$$

$$a = \frac{\frac{902{,}33}{3}}{\frac{372{,}67}{3}} = \frac{300{,}777}{124{,}223} = 2{,}42$$

$$b = 7459{,}7 - \left(2{,}42 \times 3541{,}3\right) = -1110$$

$$y = 2{,}42x - 1110$$

1 110 correspond à 31 % (0,31) du temps moyen au semi-marathon. On peut donc écrire :

$$\bar{y} = 2{,}42\bar{x} - 0{,}31\bar{x}$$

$$\bar{y} = 2{,}11\bar{x}$$

Le temps au marathon estimé à partir de cet ajustement réalisé sur trois couples de points (les performances des trois athlètes) est le double de celui du semi-marathon auquel il faut ajouter 11 % du temps moyen au semi-marathon, soit 390 secondes, soit environ 6 minutes 30.

Mesure de la qualité de l'ajustement

Pour évaluer la qualité du modèle, on calcule le coefficient de détermination ou/et le coefficient de corrélation linéaire. Lorsque l'ajustement est parfait, les valeurs de *y* estimées sont confondues avec les valeurs réelles de *y*. Plus les valeurs estimées s'écartent des valeurs réelles, moins l'ajustement est bon. Lorsque l'ajustement est parfait (les valeurs estimées sont les mêmes que les valeurs réelles), la variance des points estimés est la même que la variance des points réels. En effet, l'écart entre valeurs réelles et estimées est nul pour tous les points. Quand il existe des écarts entre valeurs réelles et estimées, la variance du modèle s'écarte de celle des points réels (variance totale). L'écart entre ces deux variances correspond à la variance des écarts entre valeurs réelles et estimées : on parle de variance résiduelle. Plus elle est grande, moins le modèle est bon.

Comprendre la démarche mais ne pas retenir

Variance résiduelle :

$$Var\,(y - \hat{y}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{1}{n}\sum_{i=1}^{n}\left[y_i - (ax_i + b)\right]^2 = \frac{1}{n}\sum_{i=1}^{n}\left[y_i - \left(ax_i + \left(\bar{y} - a\bar{x}\right)\right)\right]^2$$

$$Var\,(y - \hat{y}) = \frac{1}{n}\sum_{i=1}^{n}\left[y_i - ax_i - \bar{y} + a\bar{x}\right]^2 = \frac{1}{n}\sum_{i=1}^{n}\left[\left(y_i - \bar{y}\right) + \left(a\bar{x} - ax_i\right)\right]^2$$

$$Var\,(y-\hat{y}) = \frac{1}{n}\sum_{i=1}^{n}\left[\left(y_i-\bar{y}\right)-a\left(x_i-\bar{x}\right)\right]^2$$

$$Var\,(y-\hat{y}) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2 - 2a\frac{1}{n}\sum_{i=1}^{n}\left[\left(y_i-\bar{y}\right)\left(x_i-\bar{x}\right)\right] + a^2\frac{1}{n}\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2$$

$$Var\,(y-\hat{y}) = Var\,(y) - 2aCov\,(x,y) + a^2 Var\,(x)$$

$$Var\,(y-\hat{y}) = Var\,(y) - 2\frac{Cov(x,y)}{Var\,(x)}Cov(x,y) + \frac{Cov(x,y)^2}{Var\,(x)^2}Var\,(x)$$

$$Var\,(y-\hat{y})) = Var\,(y) - 2\frac{Cov(x,y)^2}{Var\,(x)} + \frac{Cov(x,y)^2}{Var\,(x)}$$

$$Var\,(y-\hat{y}) = Var\,(y) - \frac{Cov(x,y)^2}{Var\,(x)}$$

On montre que la variance du modèle est égale à : $\frac{Cov(x,y)^2}{Var(x)}$

$$Var\left(\hat{y}-\bar{y}\right) = Var\left(\hat{y}\right) = Var\left(ax+b\right) = Var\left(ax\right) = a^2 Var\left(x\right) = a^2 Var\left(x\right)$$

$$Var\left(\hat{y}\right) = \left[\frac{Cov(x,y)}{Var\,(x)}\right]^2 Var\left(x\right) = \frac{Cov(x,y)^2}{Var\,(x)}$$

Donc,

À retenir

Variance totale = Variance du modèle + variance résiduelle :

$$Var(y) = Var(\hat{y}) + Var(y-\hat{y})$$

$$R^2 = \frac{Var(\hat{y})}{Var(y)} = \frac{\frac{Cov(x,y)^2}{Var(x)}}{Var(y)} = \frac{Cov(x,y)^2}{Var(x)Var(y)} = \frac{\left[\frac{1}{n}\sum_{i=1}^{n}\left(x_i-\bar{x}\right)\left(y_i-\bar{y}\right)\right]^2}{\left(\frac{1}{n}\sum_{i=1}^{n}\left[\left(x_i-\bar{x}\right)^2\right]\right)\left(\frac{1}{n}\sum_{i=1}^{n}\left[\left(y_i-\bar{y}\right)^2\right]\right)}$$

$$r = \sqrt{\frac{Cov(x,y)^2}{Var(x)Var(y)}} = \frac{Cov(x,y)}{\sigma_x\sigma_y} = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(x_i-\bar{x}\right)\left(y_i-\bar{y}\right)}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left[\left(x_i-\bar{x}\right)^2\right]}\times\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left[\left(y_i-\bar{y}\right)^2\right]}}$$

Quand l'ajustement est parfait, la variance du modèle est égale à la variance totale. Dans ce cas, R^2 vaut 1. A mesure que l'ajustement est moins bon, la part de la variance résiduelle augmente et le rapport entre variances du modèle et totale diminue et tend vers 0. Donc :

$$0 \le R^2 \le 1$$

Le coefficient de corrélation linéaire varie entre – 1 et + 1. Sa valeur absolue tend vers 1 quand l'ajustement est très bon. Elle tend au contraire vers 0 quand l'ajustement est peu adapté au nuage de points. Enfin, r est négatif quand les variables évoluent en sens contraire et est positif quand les variables évoluent dans le même sens.

$$-1 \le r \le 1$$

Application numérique à l'exemple des marathoniens

Pour calculer R^2 et r, il faut simplement effectuer un calcul supplémentaire par rapport à ceux réalisés pour la détermination de la droite de régression linéaire de y en x : celui de la variance de y (tableau 4).

Tableau 4 : Organisation et détails des calculs nécessaires pour déterminer R^2 et r

Athlète	Semi (x)	Marathon (y)	x - x moy	y - y moy	(x - x moy)*(y - y moy)	(x - x moy)²	(y - y moy)²
Patrick Makam	3 532	7 418	-9,3	-41,7	387,81	86,49	1738,89
Haile Gebreselassie	3 535	7 466	-6,3	6,3	-39,69	39,69	39,69
Paul Tergat	3 557	7 495	15,7	35,3	554,21	246,49	1246,09
Somme	10 624	22 379			902,33	372,67	3024,67
Moyenne	3 541,3	7 459,7			300,777	124,223	1008,223

$$R^2 = \frac{Cov(x,y)^2}{Var(x)Var(y)} = \frac{(300{,}777)^2}{124{,}223 \times 1008{,}223} = 0{,}72$$

$$r = \sqrt{\frac{Cov(x,y)^2}{Var(x)Var(y)}} = \frac{Cov(x,y)}{\sigma_x \sigma_y} = \frac{(300{,}777)}{\sqrt{124{,}223} \times \sqrt{1008{,}223}} = 0{,}85$$

Le coefficient de corrélation est de 0,85, soit proche de 1. Il témoigne d'un bon ajustement des points par le modèle.

Ajustement d'un nuage de points par une fonction de type logarithmique

Lorsque le nuage de points ne peut être résumé par une droite de régression linéaire, il est parfois possible d'utiliser d'autres modèles. La fonction logarithmique est l'un d'eux. Mais le principe d'estimation des paramètres de la fonction est le même que celui énoncé précédemment pour la droite de régression linéaire. On utilise une « astuce » mathématique en modélisant non pas la relation entre *x* et *y*, mais celle entre le logarithme de *x* et *y* (cf. page suivante figures 9 et 10).

$$y = a\ln(x) + b$$
$$y = aX + b$$

A partir des données suivantes :

Tableau 5 : Descendance atteinte (pour 100 femmes) selon l'âge des femmes de la génération 1947

Age (x)	Descendance atteinte pour 100 femmes (y)	ln (âge) = X
15	0	2,708
20	25	2,996
25	111	3,219
30	171	3,401
35	200	3,555
40	211	3,689

vérifiez que :

$a = 242{,}0$

$b = -669{,}5$

$y = 242X - 669{,}5$

$y = 242\ln(x) - 669{,}5$

$R^2 = 0{,}96$

$r = 0{,}98$

Figure 9 : Cumul du nombre d'enfants pour 100 femmes (y) de la génération 1947 selon l'âge (x) et ajustement (en pointillés)

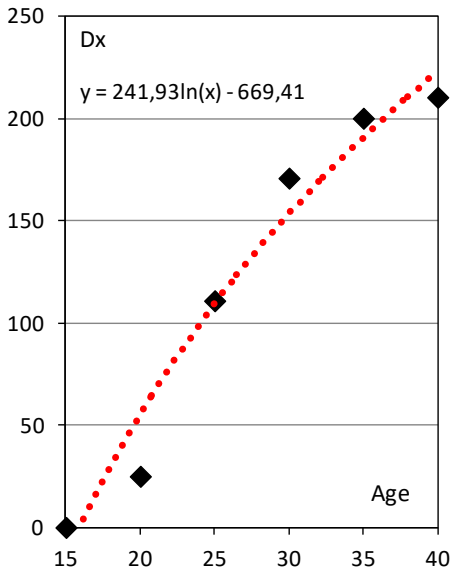
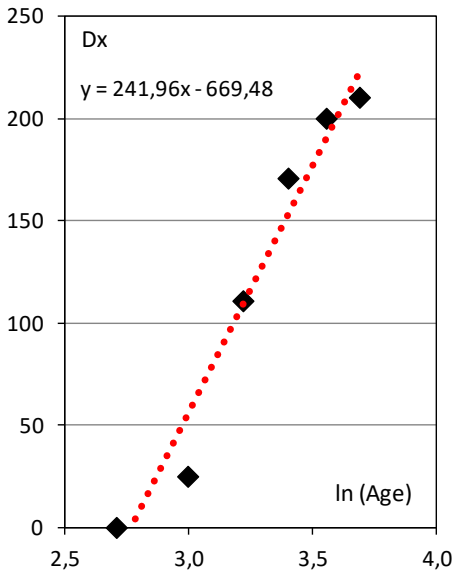


Figure 10 : Cumul du nombre d'enfants pour 100 femmes (y) de la génération 1947 selon le logarithme népérien de l'âge (ln(x)) et ajustement (en pointillés)



Ajustement d'un nuage de points par une fonction de type exponentiel (ou puissance)

Lorsque la variation de la variable *y* en fonction de *x* suit une courbe de type exponentiel, on peut ajuster *y* en fonction de *x* par un modèle exponentiel (ou puissance). En effet, comme on l'a vu lors de la présentation de la construction d'une échelle semi-logarithmique, quand *y* croît de manière exponentielle, ln(*y*) ou log(*y*) croît de manière linéaire (cf. page 13, figures 11 à 13). C'est par exemple le cas des risques de mortalité qui, à partir de l'âge de 20 ans, augmentent de façon exponentielle avec l'âge. C'est la raison pour laquelle on représente les séries de quotients de mortalité à l'aide d'une échelle semi-logarithmique[1]. Dans le cas présent, il faut donc déterminer la droite de régression linéaire de ln(*y*) en fonction de *x*.

[1] Dans le cas des risques de mortalité, on utilise le plus souvent les logarithmes en base 10. Mais on peut, comme on va le faire ici, utiliser aussi les logarithmes népériens.

Ajustement par une fonction exponentielle :

$$y = \exp(ax + b) = e^{(ax+b)} = e^{ax} \times e^{b} = e^{b} \times e^{ax} = c \times e^{ax}$$
$$\ln(y) = ax + b$$
$$Y = ax + b$$

Ajustement par une fonction puissance

$$y = 10^{(ax+b)} = 10^{ax} \times 10^{b} = 10^{b} \times 10^{ax} = k \times 10^{ax}$$
$$\log(y) = ax + b$$
$$Y = ax + b$$

A partir des données suivantes :

Tableau 6 : Risques de mortalité selon l'âge. France, 2003-2004 (hommes et femmes réunis)

Age (x)	quotients de mortalité ${}_1q_x$ (y)	$\ln({}_1q_x) = \ln(y)$ Y modèle "exponentiel"	$\log({}_1q_x) = \log(y)$ Y modèle "puissance"
20	0,00114	-6,7802	-2,9446
40	0,00564	-5,1784	-2,249
60	0,01547	-4,1691	-1,8106
70	0,03813	-3,2667	-1,4187
80	0,07859	-2,5435	-1,1046
85	0,13376	-2,0117	-0,8737
90	0,21260	-1,5483	-0,6724
95	0,29137	-1,2332	-0,5356

Vérifiez, en adoptant le principe de l'ajustement linéaire, que :

Pour la fonction exponentielle :

$$a = 0{,}073$$
$$b = -8{,}289$$
$$\ln(y) = 0{,}073x - 8{,}289$$
$$y = \exp(0{,}073x - 8{,}289) = e^{0{,}073x} \times e^{-8{,}289} = e^{-8{,}289} \times e^{0{,}073x}$$
$$R^2 = 0{,}994$$
$$r = 0{,}997$$

Pour la fonction puissance :

$$a = 0{,}032$$
$$b = -3{,}600$$
$$\log(y) = 0{,}032x - 3{,}600$$
$$y = 10^{(0{,}032x - 3{,}600)} = 10^{0{,}032x} \times 10^{-3{,}6} = 10^{-3{,}6} \times 10^{0{,}032x}$$
$$R^2 = 0{,}994$$
$$r = 0{,}997$$

Dans les deux cas, l'ajustement est excellent … mais une lecture graphique du nuage de points le montrait déjà parfaitement.

Fig. 11 : Risques de mortalité ($_1q_x$) - (y) - selon l'âge (x) et ajustement (en pointillés)

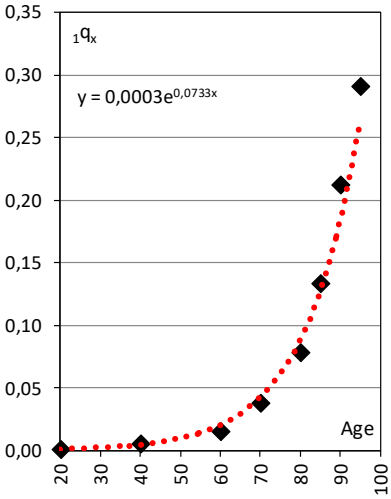


Fig. 12 : Ln ($_1q_x$) - (ln(y)) - selon l'âge (x) et ajustement (en pointillés)

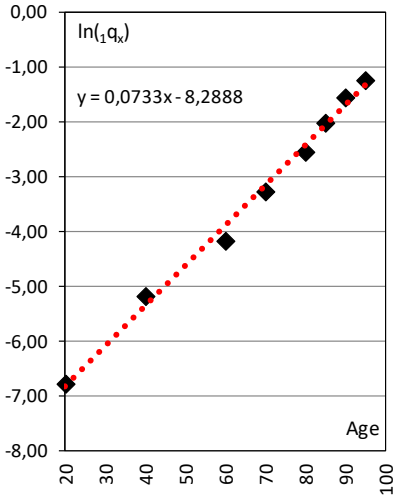


Fig. 13 : $Log_{10}(_1q_x)$ - (log(y)) - selon l'âge (x) et ajustement (en pointillés)