

# Galton et l'effet de régression

## Un peu de géométrie

Marc Bourdeau

---

Sir Francis Galton [1822-1911], on l'a vu, n'a pas compris le lien entre ce qu'il a découvert sur la tendance des descendants à revenir vers les moyennes, et la technique des moindres carrés, c'est son disciple Karl Pearson [1857-1936], qui a établi mathématiquement la chose, et donné le sens quantitatif en usage aujourd'hui au terme éponyme de *corrélation*<sup>1</sup>. Il ne faut pas négliger non plus dans ce contexte les travaux de Francis Ysidro Edgeworth [1845-1926], parent éloigné de Sir Francis, ainsi que ceux de George Udny Yule [1871-1951], tout particulièrement, pour ce dernier, en ce qui concerne les interprétations à tirer du phénomène de la régression vers la moyenne. En fait, c'est Edgeworth qui introduisit le terme de *coefficient of correlation*, un peu avant K. Pearson. Ce terme est attribué plutôt à K. Pearson<sup>2</sup>.

On présente ici les données mêmes de Galton<sup>3</sup> dont on fait l'analyse pour mettre en évidence l'*effet* dit *de régression*, l'œuvre scientifique principale de Galton. Une série d'exercices coordonnés nous mènera à la compréhension précise de ce phénomène, en réalité fort simple, qui apparaît souvent, pour ne pas dire toujours, comme très mystérieux, et, selon nous, assez pauvrement expliqué<sup>4</sup>.

---

<sup>1</sup>En anglais, on trouve souvent la référence explicite, on parle de la *Pearson correlation*, parfois de la *Pearson moment-correlation*.

<sup>2</sup>Ici encore un autre exemple de la loi d'éponymie de Stigler, dont on a parlé par ailleurs : « *No scientific discovery is named after its original discoverer.* » Alfred North Whitehead [1861-1947], celui qui a écrit les *Principia Mathematica* avec Bertrand Russell [1872-1970] a eu ce joli mot : « *Everything of importance has been said before by somebody who did not discover it.* »

<sup>3</sup>On a repris récemment les données originales pour en faire voir un tour inattendu : Amanda Wachsmuth, Leland Wilkinson, & Gerald E. Dallal (2003), *Galton's Bend : A Previously Undiscovered Nonlinearity in Galton's Family Stature Regression Data*. The American Statistician, **57**(3), 190 – 192.

<sup>4</sup>On trouve trop souvent, à l'aide d'exemples, des explications savantes pour dénoncer l'utilisation frauduleuse de l'effet de régression qui sont, même si les exemples semblent pertinents, autant d'explications ...fumeuses :  
e.g. <http://www.stat.berkeley.edu/users/stark/SticiGui/Text/ch6.htm>.

## 1 L'effet de régression

C'est en observant les tailles des pères et celles des fils que Galton a constaté que les fils de pères plus petits ou plus grands que la moyenne avaient tendance à avoir des tailles plus petites ou plus grandes que la moyenne des pères, selon le cas, mais moins loin de la moyenne que leurs ancêtres immédiats.

Il a ensuite utilisé les nombreuses données de tailles (Tab. 1) qu'il avait collectées à la fois sur les pères et les mères dont la moyenne a constitué sa base de comparaison, ainsi que leurs fils et filles, filles auxquelles il a donné un coefficient homothétique de 1,08 pour tenir compte des différences de tailles entre hommes et femmes<sup>5</sup>. Il a fait la même constatation.

Ce phénomène de *régression ou retour vers la moyenne*, on le verra dans cette suite d'exercices coordonnés, est un phénomène absolument général et tout à fait naturel dans les modèles linéaires ajustés aux moindres carrés qu'il traitait sans le savoir. Donc, rien à voir avec une propriété génétique telle que Galton l'a cru, ou plus métaphysique ainsi que beaucoup d'autres l'ont décrite dans d'autres contextes.

On trouvera les données du TAB. 1 dans le fichier qu'on peut obtenir en cliquant sur l'icone dans la marge.

1. Trouvez des estimateurs des moyennes et variances, des covariance et corrélation entre les tailles des parents et celles des enfants à partir du tableau de contingence et des valeurs moyennes des intervalles de tailles des uns et des autres. Les données de l'une et l'autre variables de taille vous semblent-elles gaussiennes ?

Nous sommes donc en présence d'un échantillon d'une bi-gaussienne dont vous avez obtenu des estimations des principaux paramètres. On vous fournit dans le même fichier de données une simulation d'une pareille bi-gaussienne des 928 sujets du tableau de contingence de Galton pour que vous puissiez les traiter à votre guise. Ces données ont été générées par le module *mvtnorm*

---

<sup>5</sup>F. Galton, *Regression towards Mediocrity in Hereditary Status*, Journal of the Anthropological Institute, **15**, 1886, 246-263. À noter qu'en anglais on trouve au moins quatre termes pour décrire le tendance centrale d'un ensemble de données : *mean*, *average*, *median*, *norm*. Le terme *mediocrity* vient du latin « *mediocritas* » comme dans « *In mediocritas stat virtus* », qui désigne le juste milieu.



Data

TAB. 1 – Dans ce tableau, on trouve les données mêmes de Galton. Les classes sont notées par les points médians des intervalles sauf pour les intervalles extrêmes.

Heights of the Mid-parents in inches.	Heights of the Adult Children.														Total Number of		Medians.
	Below	62.2	63.2	64.2	65.2	66.2	67.2	68.2	69.2	70.2	71.2	72.2	73.2	Above	Adult Children.	Mid-parents.	
Above ..	..	..	..	..	..	..	..	..	..	..	..	1	3	..	4	5	..
72.5 ..	..	..	..	..	..	..	..	1	2	1	2	7	2	4	19	6	72.2
71.5 ..	..	..	..	..	1	3	4	3	5	10	4	9	2	2	43	11	69.9
70.5 ..	1	..	1	..	1	1	3	12	18	14	7	4	3	3	68	22	69.5
69.5 ..	..	..	1	16	4	17	27	20	33	25	20	11	4	5	183	41	68.9
68.5 ..	1	..	7	11	16	25	31	34	48	21	18	4	3	..	219	49	68.2
67.5 ..	..	3	5	14	15	36	38	28	38	19	11	4	..	..	211	33	67.6
66.5 ..	..	3	3	5	2	17	17	14	13	4	..	..	..	..	78	20	67.2
65.5 ..	1	..	9	5	7	11	11	7	7	5	2	1	..	..	66	12	66.7
64.5 ..	1	1	4	4	1	5	5	..	2	..	..	..	..	..	23	5	65.8
Below ..	1	..	2	4	1	2	2	1	1	..	..	..	..	..	14	1	..
Totals ..	5	7	32	59	48	117	138	120	167	99	64	41	17	14	928	205	..
Medians ..	..	..	66.3	67.8	67.9	67.7	67.9	68.3	68.5	69.0	69.0	70.0	..	..	..	..	..

du logiciel **R**<sup>6</sup> à l'aide des commandes suivantes<sup>7</sup> :

```
> n<-928
> moy<-c(0,0)
> sig<-diag(2); diag[2,1]=diag[1,2]<- .8
> library(mvtnorm)
> x<-rmvnorm(n,mean=moy,sigma=sig)
> write.table(x,file='galton.txt', sep='\t', dec=',')
```

Notes : (1) le signe '< -' qui marque l'affectation ; (2) 'c(a,b,...)' comme dans le langage C construit des vecteurs ; (3) 'sig' est la matrice de variances-covariances ; (4) le ';' permet plusieurs commandes par ligne ; (5) la commande library(xxx) rend accessible le module ajouté xxx ; (6) la commande *rmvnorm* simule une multigaussienne (ici une bi-gaussienne) avec les paramètres indiqués ; (7) voir l'aide '?write.table' pour les détails ; (8) une fois

<sup>6</sup>On peut importer et installer le logiciel **R** en cliquant : <http://cran.r-project.org/>. On aura besoin d'installer en plus les deux modules (« packages ») *mvtnorm* (*multivariate normal*) ainsi que *ellipse*, ce dernier permettant de tracer des ellipses de confiance pour des multi-gaussiennes. Ces installations se font de façon automatique en se laissant guider après avoir cliqué l'onglet *packages* dans la fenêtre du logiciel. On ne peut vraiment se dispenser de la lecture d'un peu de documentation fournie avec le logiciel, mais nous donnons ici les commandes et fonctions nécessaires pour nos exercices.

<sup>7</sup>**R** est un langage de commandes avec le signe ">" comme *prompt*. Les valeurs des paramètres donnés ici ne sont pas ceux trouvés au numéro 1...

le fichier .txt construit, on peut l'importer dans Excel ou tout autre logiciel de traitement de données ; (9) la flèche '↑', re-redirection vers le haut, rappelle les commandes précédentes qu'on peut alors corriger et exécuter à nouveau sans les récrire au complet.

Une semblable bi-gaussienne est représentée à la Figure 1 avec son ellipse d'équiprobabilité à 95% (une partie seulement des données simulées sont représentées).

Un mot sur la construction de cette figure. La bi-gaussienne estimée pourrait se visualiser comme une cloche non circulaire mais elliptique dessinée en trois dimensions sur un plan horizontal. L'ellipse dessinée est une coupe parallèle à ce plan (une courbe de niveau en fait). Le rectangle qui borde cette ellipse est constitué des tangentes à l'ellipse, verticale et horizontale. Les dimensions du rectangle correspondent, à un facteur homothétique près (c'est le terme dû aux équiprobabilités, donc à la confiance de l'ellipse), aux écarts types respectifs des tailles des parents, en horizontal, et des enfants, en vertical. Le grand axe de l'ellipse, le segment BD, correspond au premier axe principal dans la décomposition en composantes principales, c'est la diagonale de ce rectangle, et le petit axe (non représenté) correspond au deuxième axe principal.

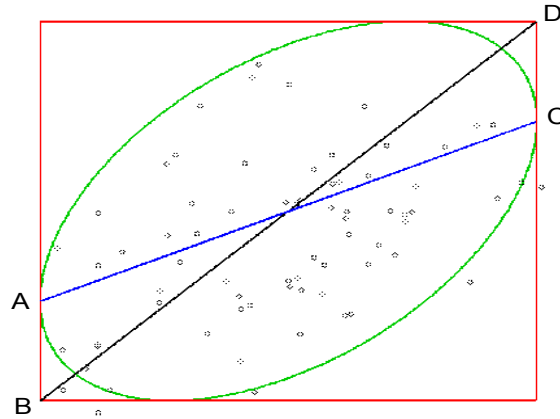


FIG. 1 – Les pseudo-valeurs des tailles des parents et des enfants. Avec l'ellipse d'équiprobabilité à 95%, les tangentes verticales et horizontales à cette ellipse. BD : le grand axe de l'ellipse ; AC : la droite des moindres carrés.

On a construit une fonction **R** qui génère la FIG. 1. Ce fichier est importé en cliquant : [l'effet régression](#). Les deux premières commandes d'exécution données plus bas rendent accessibles les modules nécessaires (qui ne sont pas accessibles par défaut même s'ils ont été installés correctement), puis

la troisième commande appelle la fonction « simulA » écrite aux fins du cours — pour permettre une certaine expérimentation. On peut en changer les paramètres par défaut qu'on peut connaître par des commandes dont un exemple est donné :

```
> library(mvtnorm)
> library(ellipse)
> sig; moy
> simulA(n,moy,sigma,conf)
```

La droite AC est la droite des moindres carrés construit sur les données. Pour comprendre cela, il faut procéder en plusieurs étapes.

2. (a) On montre d'abord que la loi marginale de  $Y|x$  est une loi gaussienne. Pour cela, il suffit, sans perte de généralité, de raisonner sur une bi-gaussienne normée avec un coefficient de corrélation  $\rho$  entre les deux termes dont la densité est :

$$f(x, y) = \frac{1}{2\pi\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}(x^2-2\rho xy+y^2)}. \quad (1)$$

Par ailleurs la marginale est évidemment

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(x-\mu)^2}, \quad (2)$$

avec, pour le cas normé :  $\mu = 0$ ,  $\sigma = 1$ . Montrez par substitution de la conjointe et de la marginale dans l'expression de la densité conditionnelle de  $Y|x$  :

$$f(y|x) = f(x, y)/f_X(x),$$

que la marginale est une gaussienne  $\mathcal{N}(\rho x; 1 - \rho^2)$ . Il devrait vous apparaître étonnant — voir la FIG. 1 —, que sa variance ne change pas selon la valeur de  $x$  ? Explications requises...

- (b) En déduire à partir du concept d'ellipse de confiance que les points milieux des segments verticaux constituent un segment AC sont les points de maximum des densités marginales  $Y|x$ , et donc la droite des moindres carrés de la régression de  $Y$  sur  $X$ .
- (c) Notons enfin que le segment AC qui passe par les points milieux de verticales coupe justement les droites verticales de tangence aux points de tangence de l'ellipse. Pour le comprendre, il faut raisonner sur les limites lorsqu'on se déplace le long de la droite AC vers les extrémités.

- (d) Montrez que les deux segments AC et BD se coupent au centre de gravité des données  $(\bar{x}, \bar{y})$ .
- (e) Dédurre aussi de la question (2a) que la droite des moindres carrés calculée sur des données normées (centrées et réduites) pour une modèle simple est de pente  $\rho$ , la corrélation entre les deux variables.
- (f) Pour bien river le clou sur les moindres carrés, montrez que pour des données quelconques  $x_i, i = 1, \dots, n$  l'expression suivante est minimale pour  $c = \bar{x}$ , et donnez-en l'interprétation dans ce contexte :

$$\sum_1^n (x_i - c)^2.$$

- (g) On pourrait construire aussi l'autre droite qui joint les deux autres points de contact de l'ellipse avec le rectangle tangent. Déterminez la signification statistique de cette autre droite, compte tenu du sens de AC.

Maintenant, raisonnons à partir de la FIG. 1 sur les propriétés statistiques de cette ellipse.

3. (a) Que se passe-t-il lorsque, pour une corrélation fixée, les écarts types des deux variables se rapprochent l'un de l'autre ? S'éloignent l'un de l'autre.
- (b) Si, à l'inverse, pour des écarts types fixés, la corrélation entre les deux variables tend vers 1 ? Vers 0 ?
- (c) Quelle est le plus grand angle possible entre les segments AC et BD ?

On est en mesure maintenant de décrire précisément l'*effet de régression* qui a tant intrigué Galton, et d'en fixer les limites.

Galton a énoncé<sup>8</sup> le principe suivant : « *It is a universal rule that the unknown kinsman in any degree of any specified man, is probably more mediocre than he.* » Par cela il signifiait que si les parents sont éloignés de la

---

<sup>8</sup>F. Galton, *op. cit.*. Dans cet article il parlait spécifiquement des tailles, mais son concept était très général, et s'appliquait tout autant à des qualités mentales que physiques, et cela pour toutes les espèces animales ! C'est à cause des difficultés de mesure pour les qualités mentales qu'il s'est limité aux qualités physiques... Dans la même veine, il a milité pour les mariages arrangés en vue de la *race betterment*, sans doute pour contrer les effets de la régression vers la médiocrité...

moyenne pour une certaine mesure, les descendants ont tendance eux-aussi à être éloignés de ladite moyenne, mais moins éloignés que leurs parents. C'est cette tendance au retour vers la moyenne qu'il appela l'*effet de régression*<sup>9</sup>.

Il illustra son article par une figure qu'il a construite sur ses données et que nous reproduisons ici (FIG. 2, en haut).

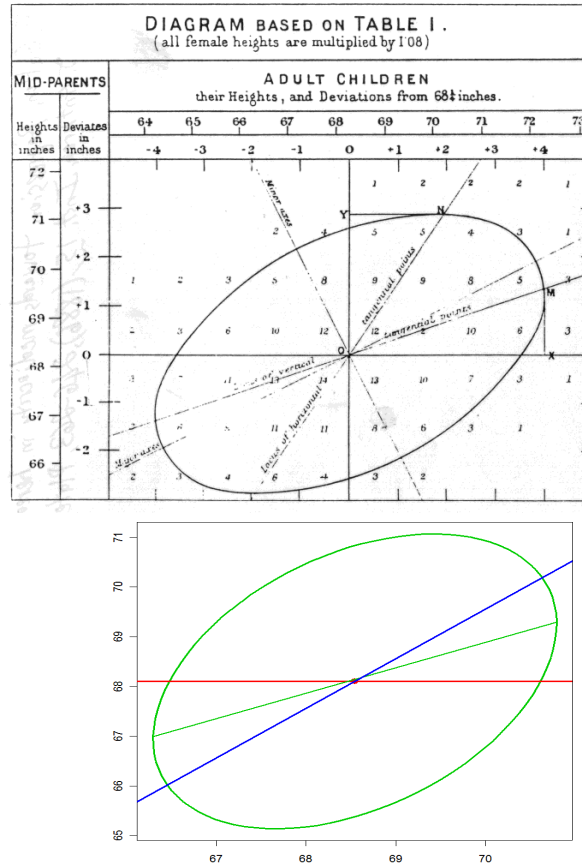


FIG. 2 – En haut, la figure de l'article de Galton qu'il a tirée de son Tab. 1. On y reconnaît les éléments géométriques de la Fig. 1. En bas, les éléments géométriques qui permettent de montrer l'effet de régression dans ce cas particulier. En rouge la droite  $y = \bar{y}$ ; en vert la droite de la régression; en bleu la droite de pente 1 passant par  $(\bar{x}, \bar{y})$ . L'ellipse de confiance des données simulées qui ont servi est de niveau  $1 - \alpha = 0,50$ .

Voyons ce qu'il en est de l'*effet de régression* dans ce cas particulier. Extrayons d'abord de la FIG. 1 les principaux éléments pour notre démonstration de l'effet de régression, tout en y ajoutant un élément crucial que

<sup>9</sup>Du latin *regressio*, retour.

Galton a bien sûr représenté sur sa figure : il s'agit de la droite horizontale qui passe par la moyenne des tailles des enfants.

On trouve au bas de la FIG. 2, construite sur des données simulées ayant les mêmes caractéristiques que celle des Galton trouvées au numéro 1, les éléments suivants<sup>10</sup> : (1) en rouge la valeur moyenne  $\bar{y} \doteq \bar{x}$  de la taille des enfants, ainsi que le centre de gravité du nuage des points *Tailles des parents*  $\times$  *Tailles des enfants*,  $(x_i, y_i)$  (non représentés), dont les moyennes sont identiques à des poussières près ; (2) en vert l'ellipse d'équiprobabilité à 50%, ainsi que la droite de régression ; (3) en bleu la droite de pente 1 qui passe par le centre de gravité du nuage des points.

Ainsi, puisque la droite en bleu est de pente unitaire, une « diagonale » donc qui passe par le centre de gravité du nuage des points, si on considère un point  $x$  de la taille des parents éloigné de la moyenne au-delà ou en-deçà qu'on notera sur la droite en rouge du graphique au bas de la FIG. 2, la perpendiculaire menée à la droite en bleu — la « diagonale » — la coupe en un point et forme un segment de même longueur que celui qui lie la valeur de l'abscisse au centre de gravité, et on constate que la droite de régression en vert est plus près de la moyenne que ne l'est la taille des parents, puisque la droite de la régression est au-dessus de la « diagonale » à gauche de la moyenne et inversement pour les abscisses à droite de la moyenne : on a  $|\hat{y} - \bar{y}| < |x - \bar{x}|$ . Voilà bien la tendance pour les enfants d'être plus près de la moyenne que pour les parents qui en sont éloignés. Cela est d'ailleurs vrai pour toute valeur de la taille des parents qui n'est pas exactement sur la moyenne, pas seulement ceux qui sont « éloignés ».

Cela tient évidemment à ce que les unités des «  $x$  » et des «  $y$  » sont les mêmes, et donc les distances sont comparables d'un axe à l'autre ayant les mêmes unités. C'est ainsi, à l'aide de ces éléments graphiques, que l'hérédité des enfants transmise par les parents, la sémantique du modèle, a incité Galton au terme faisant appel au « retour ».

Que les moyennes des «  $x$  » aient été égales ou non à celles des «  $y$  » n'aurait rien changé : l'effet de régression eût été encore présent, ce qui est démontré par la construction géométrique elle-même.

Mais n'avons-nous pas là tout de même quelque cas particulier ? Il est évident que l'effet de régression est indépendant des moyennes des variables, prédictive et de la réponse. Ce qui constitue l'effet, c'est que la droite de régression est située entre la droite moyenne «  $y = \bar{y}$  », et la droite de pente unitaire qui passe par le centre de gravité  $(\bar{x}, \bar{y})$ . Cela serait-il toujours le cas ? Pour explorer ce point, nous utilisons à nouveau le fichier R dont une

<sup>10</sup>On pourra se poser des questions sur les différences fines entre les deux parties de la figure.



fonction a créé l'image présentée ici.

Plus bas les commandes **R** qui permettent de créer par simulation d'autres graphiques tels celui du bas de la FIG. 2 à partir du fichier **R** déjà importé. Les deux premières rendent accessibles les modules nécessaires (qui ne sont pas accessibles par défaut), puis la troisième commande appelle la fonction « effetR » écrite aux fins du cours :

```
> library(mvtnorm)
> library(ellipse)
> effetR(n,moy,sigma,conf)
```

où  $n$  est le nombre de points de la bi-gaussienne simulée,  $moy$  sa moyenne dans le plan,  $sigma$  sa matrice de variances-covariances, et  $conf$  la confiance demandée pour l'ellipse. Les paramètres de la fonction sont fixés par défaut et modifiables à volonté. Ce qui est crucial, ce sont évidemment les valeurs de la matrice des variances-covariances. Par exemple, en changeant la variance des  $x_i$  on change leur dispersion, mais pas seulement... Ainsi les commandes :

```
> sigma[1,1]<- 1
> effetR(n,moy,sigma,conf)
```

viennent troubler quelque peu l'effet de régression qu'il faut bien se garder de croire universel ! La FIG. 3 montre géométriquement l'absence du phénomène de régression que des développements mathématiques doivent confirmer.

Divers essais avec le logiciel permettent de subodorer que lorsque les variances sont voisines, on trouve bien l'effet de régression, auquel cas il est tout à fait naturel (au sens statistique du terme), et ne vient aucunement de l'application elle-même, ne correspond en aucun cas donc à une interprétation « métaphysique »...

Pour mieux comprendre le phénomène, complétons la suite des exercices précédents.

Vous aurez besoin de la formule suivante pour la densité d'une bi-gaussienne générale,  $(X, Y)$ , de moyennes  $\mu_1, \mu_2$ , de variances  $\sigma_1^2, \sigma_2^2$ , et de corrélation  $\rho$  :

$$f(x_1, x_2) = \frac{1}{2\pi \sigma_1 \sigma_2 \sqrt{1 - \rho^2}} \exp \left\{ -\frac{1}{2(1 - \rho^2)} \left[ \left( \frac{x_1 - \mu_1}{\sigma_1} \right)^2 - 2\rho \frac{(x_1 - \mu_1)(x_2 - \mu_2)}{\sigma_1 \sigma_2} + \left( \frac{x_2 - \mu_2}{\sigma_2} \right)^2 \right] \right\} . \quad (3)$$

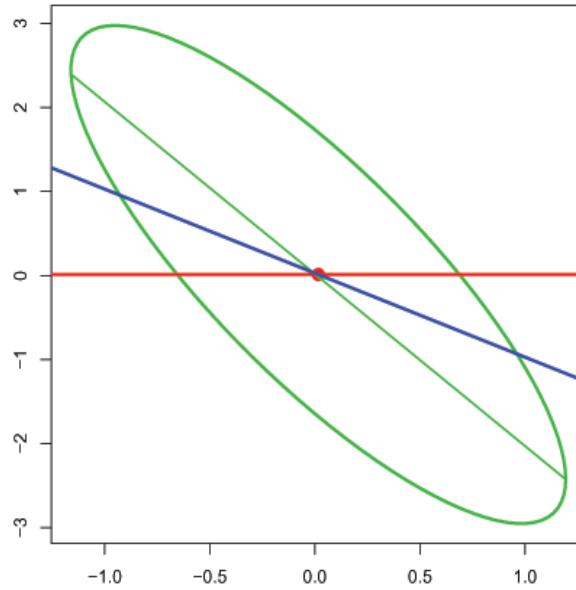


FIG. 3 – Pas d'effet de régression lorsque  $\sigma_X^2 = 1$ ,  $\sigma_Y^2 = 6.34$  et  $\sigma_{XY} = -2.06$ .

4. (a) Expliquez pourquoi l'effet de régression est indépendant des moyennes respectives des deux variables, et qu'on peut donc raisonner sans perte de généralité sur des moyennes nulles pour l'une et l'autre.
- (b) Calculez à partir des formules (2) et (3) dans ce cas spécifique, avec cependant des variances  $\sigma_X^2$ ,  $\sigma_Y^2$  différentes l'expression de la densité conditionnelle

$$f(y|x) = \frac{f(x, y)}{f_X(x)}.$$

- (c) En déduire que cette loi conditionnelle est une gaussienne, nommément :

$$Y_{|X=x} \sim \mathcal{N}\left(\frac{\sigma_Y}{\sigma_X}\rho x; \sigma_Y^2(1 - \rho^2)\right). \quad (4)$$

- (d) Montrez que l'effet de régression peut se lire de la façon suivante :

$$|E(Y|X = x)| \leq |x|$$

À noter que la variance de  $Y_{|X=x}$  est inférieure à celle de  $Y$ .

- (e) Donner maintenant la condition générale qui lie les variances et la corrélation à l'effet de régression.

- (f) Montrez que dans le cas particulier où les écarts types de  $X$  et  $Y$  sont égaux, l'effet de régression est toujours présent.
- (g) Illustrez par un graphique dans le plan  $(\rho, \frac{\sigma_Y}{\sigma_X})$  la région où l'effet de régression s'applique.
- (h) Approximativement jusqu'à quelle limite pour l'écart type des tailles des enfants, Galton aurait-il pu observer son phénomène de régression ?

On a vu au numéro 2g le sens à donner à la droite qui lie les deux autres points de contact de l'ellipse avec le rectangle tangent. Au vu de ce qu'on vient de mettre en évidence, on peut formuler un complément à l'effet de régression de  $Y$  sur  $X$ .

5. (a) Sous les mêmes hypothèses qu'au numéro 4a, vous pouvez montrer sans calculer l'analogue de la formule (4) :

$$X_{|Y=y} \sim \mathcal{N}\left(\frac{\sigma_X}{\sigma_X} \rho y; \sigma_X^2(1 - \rho^2)\right). \quad (5)$$

- (b) Donnez maintenant une interprétation de cet effet.
- (c) Donnez la condition générale qui lie les variances et la corrélation pour qu'on ait

$$|E(X|Y = y)| \leq |y|,$$

et donnez une interprétation de cette inégalité en termes d'un autre effet de régression que celui noté à l'équation (4).

- (d) On a un *double effet de régression* quand les deux conditions générales sont satisfaites. Obtenez la région du plan  $(\rho, \frac{\sigma_Y}{\sigma_X})$  qui satisfait simultanément aux deux conditions.

Peut-on trouver des paramètres où un seul des deux effets est présent mais pas l'autre ? Illustrez par des exemples les situations où un trouve un seul effet de régression mais pas l'autre, de même qu'un autre où on trouve les deux, et un autre où on ne trouve aucun des deux.

On pourra utiliser pour les illustrations la fonction **R** du fichier déjà importé : *DBeffetR(n,moy,sig,conf)*, avec les paramètres de moyennes, *moy*, et la matrice de variances-covariances, *sig*, appropriés.

Il est intéressant de mentionner ici, le moment s'y prête, une petite difficulté logique des modélisations statistiques. L'exemple suivant est bien connu<sup>11</sup>. Les couples hommes-femmes ont tendance à se former selon des affinités mentales voisines. C'est ainsi qu'on a noté une corrélation  $\rho = 0,7$  entre les quotients intellectuels (QI) des hommes et des femmes en situation de couple. Une étude statistique plus fine a montré que le modèle linéaire qui lie la valeur du QI de l'homme à celui de sa conjointe est satisfaisant.

6. (a) En utilisant le fait que le QI est une mesure de moyenne typiquement de 100 avec un écart type de 15. Observe-t-on dans ce cas un double effet de régression ?

Les équations (4) et (5) permettent de répondre aux questions suivantes en utilisant les transformations qui centrent les variables aléatoires (pour une VA quelconque  $X$  :  $X' = X - \mu$  est une variable de moyenne nulle et de même variance que  $X$ ), ou encore en raisonnant les distances en termes de proportion d'écart type à la moyenne.

- (b) Quelle est la prédiction du QI d'un homme dont la conjointe a un QI de 140 ?
- (c) Quelle est la prédiction du QI de la conjointe d'un homme dont le QI=128 ?
- (d) Appliquez cette démarche prédictive en cascade aux données de Galton.
- (e) Vous devriez être un peu perplexe devant ces prédictions ? Comment sortir de cette aporie ?

Autre approche. On peut traiter des considérations précédentes par un modèle un peu différent du modèle linéaire simple pour lier les deux variables  $Y$  et  $X$ . Pensons à la sémantique suivante :  $X$  est une caractéristique d'un individu pris au hasard dans une certaine population, une réponse à un test par exemple, mais cette réponse est sujette à un aléa, par exemple le sujet est particulièrement tendu, ou chanceux, ou malchanceux. On suppose donc l'existence d'une autre variable aléatoire,  $Z$ , modélisant cet aléa supplémentaire. Ainsi  $Y$ , la réponse observée peut s'écrire :

$$Y = X + Z \quad (6)$$

Supposons aussi que  $X$  et  $Z$  soient indépendantes, que  $Z \sim \mathcal{N}(0; \sigma_Z^2)$ , et que  $X \sim \mathcal{N}(0; \sigma_X^2)$ . Comme plus haut, on se ramène sans perte de généralité à des variables centrées.

---

<sup>11</sup>Voir <http://www.stat.berkeley.edu/users/stark/SticiGui/Text/ch6.htm>.

7. (a) Expliquez la différence entre un modèle linéaire simple de régression de  $Y$  sur  $X$  et celui-ci.
- (b) Montrez que  $Y \sim \mathcal{N}(0; \sigma_Z^2 + \sigma_X^2)$
- (c) Montrez (ce qui est presque évident par définition de  $Y$ ) que l'espérance conditionnelle de  $Y$  étant donné  $X = x$  vaut tout simplement  $x$  :  $E(Y | X = x) = x$ . Donnez une interprétation en termes de la sémantique énoncée ci-dessus.
- (d) Montrer que la corrélation entre  $X$  et  $Y$  vaut :

$$\rho_{XY} = \frac{\sigma_X}{\sqrt{\sigma_Z^2 + \sigma_X^2}}.$$

- (e) Trouvez, en adaptant la réponse de l'exercice 4b, l'expression pour la densité conditionnelle :

$$f(x|y) = \frac{f(x, y)}{f_Y(y)}.$$

- (f) En déduire que :

$$E(X | Y = y) = \frac{\sigma_X^2}{\sigma_Z^2 + \sigma_X^2} y,$$

plus proche que  $y$  de la moyenne de  $X$  qui est 0. Donnez une interprétation en terme du phénomène de régression.

Nous avons donc une dérive importante du phénomène de régression. On a donc On montre cependant que lorsqu'on restore l'égalité des variances de  $X$  et de  $Y$  par une transformation appropriée de  $Y$ , on retombe sur ce qu'on connaît.

- (g) En changeant l'échelle de  $Y$  de sorte que sa variance devient celle de  $X$ , par l'homothétie :  $Y' = \rho Y$ , on rétablit alors l'égalité  $\sigma^2$  des variances de  $X$  et de  $Y$ . Montrer alors que

$$E(Y' | X = x) = \rho x \text{ et } E(X | Y' = y) = \rho y.$$

De plus on a :

$$V(Y' | X = x) = V(X | Y' = y) = (1 - \rho^2)\sigma^2.$$

Ainsi donc, en modifiant un peu la notation précédente : pour des variables  $X, Y$ , ayant mêmes variances  $\sigma^2$  et une corrélation  $\rho^2$ , on a le double effet de régression :

$$E(Y | X = x) = \rho x \text{ et } E(X | Y = y) = \rho y.$$

$$V(Y | X = x) = V(X | Y = y) = (1 - \rho^2)\sigma^2.$$

Bien comprendre pourquoi ces expressions expliquent le terme de *double effet de régression vers la moyenne*.

- (h) En vous plaçant dans la situation concrète de la cueillette de données selon le modèle (6), commentez-en la pratique.

## 1.1 L'effet de ...Galton

Ainsi donc, dès qu'on quitte ce cas particulier des variables ayant mêmes variances, l'effet de régression n'est pas toujours avéré, beaucoup s'en faut. De surcroît, dans le cas particulier où il l'est, l'effet de régression n'est pas du tout lié à une quelconque propriété provenant de l'application elle-même, à sa sémantique propre, mais est un fait mathématique (statistique) toujours vrai.

Après Galton, et ce très rapidement, on a étendu le terme de régression à la méthode elle-même de détermination des paramètres des modèles linéaires par moindres carrés<sup>12</sup>. Et l'effet de régression fut étendu à tout modèle, chargé *de facto* des interprétations les plus farfelues liées à sa sémantique, confortant chez beaucoup l'adage qu'*on fait dire n'importe quoi aux statistiques*<sup>13</sup>.

Même si l'utilisation frauduleuse quoique souvent involontaire de l'effet de régression fut dénoncée à maintes reprises, mais pas toujours très clairement, les mathématiques n'étant pas ce qu'on pourrait appeler évidentes, il en reste toujours quelque chose, tant il est vrai que la nature humaine a pour première propriété de vouloir donner du sens au monde environnant...

### Définition.

*L'erreur de régression* consiste à attribuer à une cause externe ce qui est purement un effet statistique.

Ainsi on remarque dans les cas de test-retest, que les personnes ayant particulièrement bien réussi au test, sont moins bonnes au retest et inversement. Rien d'autre là que l'effet de régression, puisqu'en effet on note la plupart du temps que dans les cas de test-retest, on trouve des moyennes et des variances voisines.

<sup>12</sup>Voir à cet égard l'excellent livre : Stephen M. Stigler, *Statistics on the Table*, Cambridge University Press, Cambridge MA, 1999.

<sup>13</sup>Milton Friedman, *Do old fallacies ever die ?*, Journal of Economic Literature, **30**, 1992, 2129-2132. Milton Friedman fut le récipiendaire du Prix Nobel d'économie en 1976.

Dans le site référé plus ci-dessus, on trouve l'exemple suivant typique de ce genre d'erreur d'appréciation<sup>14</sup>. On a fait une étude dans l'aviation israélienne pour déterminer les effets des récompenses-réprimandes. Ainsi, les pilotes ayant fait un atterrissage particulièrement réussi étaient cités en exemple, réprimandés dans le cas contraire. On a observé qu'au vol suivant, les réprimandés avaient tendance à mieux réussir mieux leur atterrissage, les loués moins bien. Cela contredisait bien sûr la théorie qui veut qu'une rétroaction *positive* soit plus efficace qu'une *négative* pour susciter des améliorations (tous les parents savent cela!), ce qui valut un article aux auteurs... Mais il s'agissait là seulement de l'effet de régression ! De l'effet à l'erreur de régression...

Galton a eu une nombreuse descendance dont les « tailles » ne furent pas toujours à sa hauteur ! Et même, pour revenir à son exemple paradigmatique, la taille des générations successives, il importe de noter qu'il faut postuler une invariance des variances des tailles humaines au cours des générations. On sait bien à quel point la nourriture, la consommation de calcium en bas âge entre autres, est un déterminant essentiel de la taille des individus. Ainsi, on a vu dans l'après-guerre<sup>15</sup> une augmentation significative de la taille des individus. Il suffirait alors, on l'a compris, que la dispersion de la taille des enfants soit suffisamment plus grande que celle des parents pour qu'on ait le contraire de la régression vers la moyenne... D'ailleurs dans l'esprit de Galton, les tailles moyennes des enfants et des parents était très voisines, de même que leurs variances.

À noter que même dans le cas de Galton, l'effet de régression à long terme eût dû avoir pour ...effet de réduire les tailles à une taille unique. Et si le passé était garant du futur, il eût pu être étonné d'une diversité rémanente des tailles humaines.

---

<sup>14</sup>Dans <http://www.stat.berkeley.edu/users/stark/SticiGui/Text/ch6.htm>, on rapporte les résultats de l'article suivant : Tversky & Kahnemann, *Judgement under Uncertainty : Heuristics and Biases*, Science **185**, 1974, pp1124-1131.

<sup>15</sup>On parle encore ainsi de la période du siècle dernier qui a suivi la Deuxième guerre mondiale de 1939 à 1945.

## 2 Autres explications géométriques

On ne trouvera pas dans ce cours d'applications des considérations géométriques qui vont suivre, mais tout de même, il peut être intéressant d'avoir sous les *yeux de l'esprit* ces éléments de théorie, d'autant plus que les moyens d'observations des données sont devenus de plus en plus faciles à mettre en œuvre.

Pour l'illustration des éléments géométriques, on se reportera à la FIG. 4 construite sur les données de Galton<sup>16</sup>. Quand on représente dans un plan cartésien les données  $(x_i, y_i)$  pour lesquelles un modèle simple semble assez juste, on voit immédiatement, on la dessine mentalement, une certaine ellipse de confiance qui résume, en quelque sorte, les données (ellipse de confiance à 90% ici). De là on reconstitue aisément le rectangle  $FGKL$ , ainsi que la droite des moindres carrés  $AC$ .

Le triangle  $ABC$  est construit de façon évidente. Quant à la droite  $DE$ , elle est la verticale la plus longue intérieure à l'ellipse : elle passe par le centre de gravité du nuage de points,  $(\bar{x}, \bar{y})$ .

Les longueurs des droites  $\overline{BC}$ ,  $\overline{DE}$ , ainsi que  $\overline{FG} = \overline{KL}$  changent proportionnellement selon la confiance de l'ellipse, toutes autres choses étant égales par ailleurs, mais, dans tous les cas, les propriétés suivantes (données ici sans preuve) sont vraies :

- La corrélation entre les deux variables est le rapport :  $\overline{BC}/\overline{FG}$ .
- $(\overline{BC})^2 + (\overline{DE})^2 = (\overline{FG})^2$ . C'est l'équation de l'analyse de la variance du modèle.
- Comme pour cette équation, le rapport  $(\overline{BC}/\overline{FG})^2$  représente la fraction de la variation totale présente dans les données expliquées par le modèle.
- Et le rapport  $(\overline{DE}/\overline{FG})^2$  représente la part de la variation non expliquée par le modèle.

Pour mieux comprendre ces éléments visuels, on se reportera aux deux cas de figure de la FIG. 5, dont les corrélations sont respectivement de -0,90 à gauche et de 0,10 à droite.

---

<sup>16</sup>Et pour plus de détails au texte : George W. Cobb, Jeffrey A. Witmer & Jonathan D. Cryer, *An electronic companion to Statistics*, Cogito Learning Media Inc., New York, 1997.



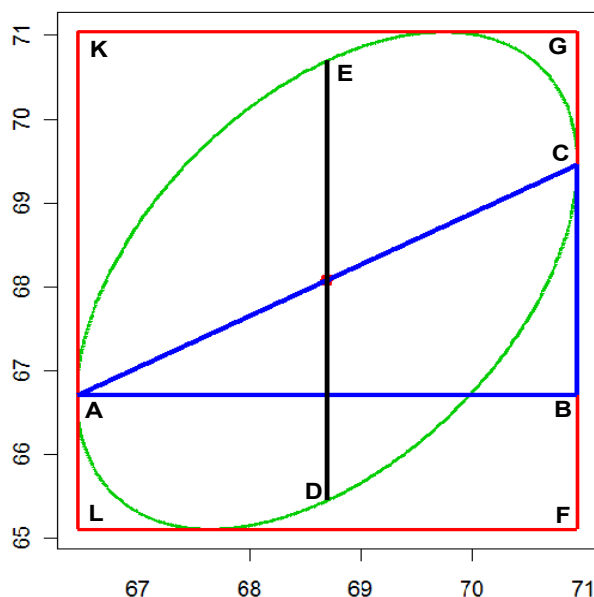


FIG. 4 – Éléments géométriques significatifs déterminés par l'ellipse de confiance en vert. Construits sur les données de Galton de corrélation 0,46.

On pourra expérimenter avec ces éléments en utilisant la fonction *ellipsePlus* dans le programme **R** invoqué plus haut. Les valeurs de  $n$  (le nombre de points de la bi-gaussienne à simuler),  $moy$  (le vecteur des deux moyennes),  $sig$  (une matrice de variances-covariances qu'on peut transformer en matrice de corrélations) et  $conf$  (la confiance des ellipses) sont fixées par défaut, mais sont modifiables à volonté, et la fonction à appeler est *ellipsePlus* avec ses paramètres : l'exemple suivant change les moyennes, et demande une corrélation de -0,8.

```
> library(mvtnorm); library(ellipse)
> conf<-0.75
> moy<-c(68.31, 68.09)
> sig[1,2]=sig[2,1]<- -0.8
> ellipsePlus(n,moy,sigma,conf)
```

Le résultat est un vecteur de composantes : (1)  $\overline{DE}^2$ , (2)  $\overline{BC}^2$  ainsi que (3)  $\overline{DE}^2 + \overline{BC}^2$  qu'il faut comparer à (4)  $\overline{FG}^2$ . Attention, on a calculé le vecteur  $\overline{DE}$  par interpolation linéaire, ce qui donne une légère différence entre les deux derniers résultats présumément égaux.

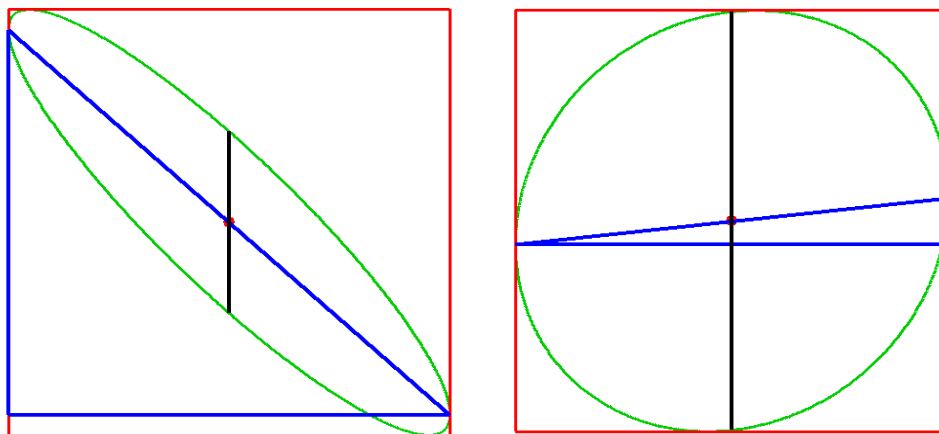


FIG. 5 – Deux cas opposés des éléments visuels de la régression : à gauche une corrélation entre les  $x_i$  et les  $y_i$  de  $-0,90$ , à droite de  $0,10$ .

8. En aucune façon ces vérifications empiriques ne constituent une preuve (voir la FIG. 4 ) que

$$(\overline{BC})^2 + (\overline{DE})^2 = (\overline{FG})^2.$$

On donne une médaille en chocolat à la première personne qui trouve une preuve de ces faits, et deux médailles en chocolat (et du bon !) à la première personne qui trouve la preuve par une astucieuse construction géométrique.

**Remerciements à André Giroux.** Le professeur Giroux, du département de Math & Stat de l'Université de Montréal a donné une preuve en utilisant Mathematica. Il s'est mérité une médaille en chocolat, avec les remerciements de l'auteur des ces lignes. Qui se fera un plaisir de l'envoyer — la preuve ! — à quiconque la demandera.