Le lien entre deux variables quantitatives

Bernard BRANGER
Epi-Sûre
(Association de conseil en épidémiologie et statistiques)
11 bis, rue Gabriel Luneau - 44000 NANTES
Mail : branger44@gmail.com - Tél 06 32 70 33 80

1. Deux méthodes possibles

- Si les variables sont « symétriques » dans leur influence biologique (il ne s'agit pas de symétrie statistique), l'une n'expliquant pas l'autre, la méthode est **le coefficient r de corrélation de Pearson ou ρ de Spearman.**

- Si une variable explique l'autre, l'une est dite dépendante (x) ou à expliquer (la maladie), et l'autre est dite explicative (y). **La méthode est la régression linéaire,** et l'indice est la pente de la droite (test par rapport à 0).

2. Deux précautions

- Si les variables sont **« trop proches »** (l'une est comprise dans l'autre par exemple), la corrélation est trop évidente ; alors, r ou pente sont souvent significatifs, mais c'est de la « tautologie » ou « pseudo-évidence »..

- Si les variable sont « **trop lointaines** » (pas de lien évident entre les deux), et si r ou pente sont significatifs, çà peut n'avoir aucun sens : deux variables peuvent évoiuer en même tems sans qu'il y ait de lien causal. Par exemple, en France, le chômage croit avec la baisse de la pratique réligieuse au fil des années !

3. Exemple : Corrélation entre des séries de 25 nombres au hasard et leurs différences

Nbres A	Nbres B	Différence D
11	50	39
47	64	17
21	96	75
59	96	37
44	59	15
15	21	6
58	23	-35
99	95	-4
20	53	33
75	96	21
75	51	-24
91	25	-66
16	92	76
30	12	-18
63	40	-23
43	40	-3
61	4	-57
68	93	25
74	7	-67
54	89	35
1	67	66
2	26	24
59	99	40
20	74	54
65	75	10
49	45	-4

D = B - A

	r	
Corr A-B	0,081	Pas de corrélation entre deux séries de nombres au hasard
Corr A-D	-0,62	Corrélation entre deux séries de nombres liées entre elles
Corr B-D	0,73	

4. Régression vers la moyenne

Extraits : La politique des grands nombres. Alain Desrosières. La Découverte. 1993

Galton (Francis, cousin de Darwin, anglais, 1822 - 1911) décide, en 1885, de mesurer en Angleterre la taille de personnes volontaires en la comparant à la taille de leurs parents. Remarquons que ceci se déroule lors d'une exposition universelle, et que ce sont les individus qui financent cette recherche...

Nous raisonnerons sur la taille des pères et sur la taille des fils. Galton observe que, EN MOYENNE, la taille des enfants est la même que celle de leur parents, et que, donc, EN MOYENNE, que les parents grands ont des enfants grands. De plus, il observe que la dispersion (c'est-à-dire la variance) est la même pour les pères que pour les fils (sy = sx) : il n'y donc pas de variance propre aux fils, qui s'ajouterait à la variance des pères.

Il dresse alors un plan à deux axes : taille du père en abscisse (contrôlée), et taille du fils (à expliquer, aléatoire). La droite de régression passe par le point (moyenne des fils - moyenne des pères), et comme coefficient de corrélation r < 1, la pente de la droite vaut $\mathbf{p_o = r\, s_y/s_x = r < 1}$. Dans l'expérience de Galton, elle vaut 2/3 = 0.66.

Sur la figure, on constate donc que lorsque $x_i > m_x$ (taille des pères grande), $y_i < y'_i$, c'est-à-dire que les enfants des pères grands sont, EN MOYENNE, plus petits que leurs pères. Il en est de même pour les enfants des pères petits qui sont, EN MOYENNE, plus grands que leurs pères. Cela explique le fait que les variances pères et fils soient égales : il existe un « tassement » autour de la moyenne. Le **terme de régression**, que Galton a inventé, s'applique ici tout à fait, et c'est ce nom-là qui sert aujourd'hui pour parler

du lien entre deux variables quantitatives. Le terme régression « vers la moyenne » en est le corollaire quand on compare les tailles extrêmes des pères. Deux commentaires :

1. Les conclusions de cette observation se sont intégrées dans les discussions scientifiques, philosophiques et politiques de l'époque : pour ceux qui comparent la pente à 0 (cas scientifique habituel), la droite de régression confirme le concept d'hérédité des caractères innés (quand les pères sont grands, les enfants sont grands). C'est le courant de Darwin. D'autres comparaient, en fait, la pente à 1 et voulaient mettre en évidence le terme de régression au sens littéral et social : la « race » s'appauvrit en régressant vers la moyenne !! C'est Cheysson (Emile, français, 1836 - 1910) et Durkeim (Emile, français, 1858 - 1917) qui utiliseront ce concept.

2. Dans notre pratique quotidienne, le phénomène « régression vers la moyenne » nous guette en permanence (J Martin Bland, Douglas G Altman. Regression towards the mean, BMJ 1994; 308: 1499, et BMJ 1994; 309 : 780)
   - comparaisons de deux mesures uniquement sur les résultats pathologiques : mesure de la tension artérielle chez les hypertendus avant et après traitement, ou mesure du cholestérol. Nécessité pour contrer la phénomène « régression » d'une groupe contrôle.
   - décisions et pratiques cliniques
   - double mesure d'un paramètre et pente différente de 1
   - biais de publications !

Figure : Modèle théorique du lien taille du père – taille du fils avec une pente à 0.66

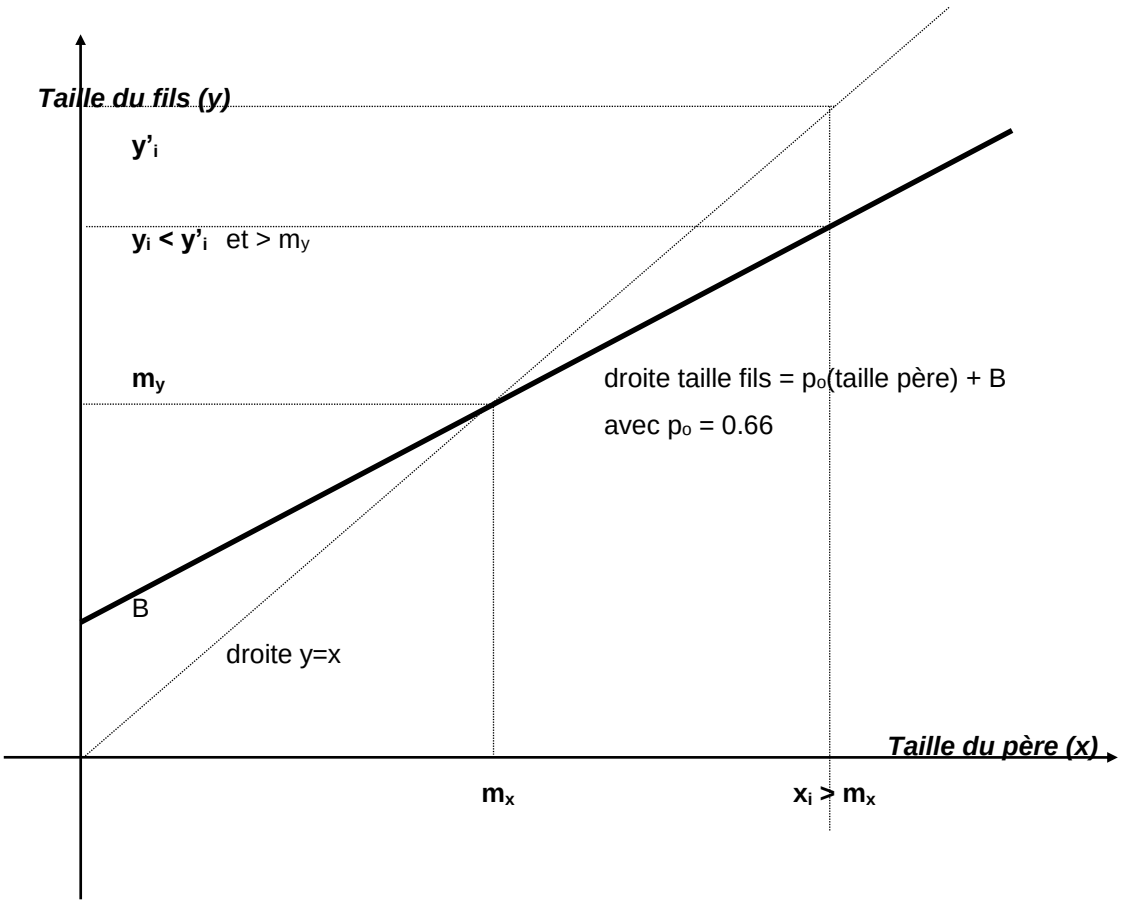
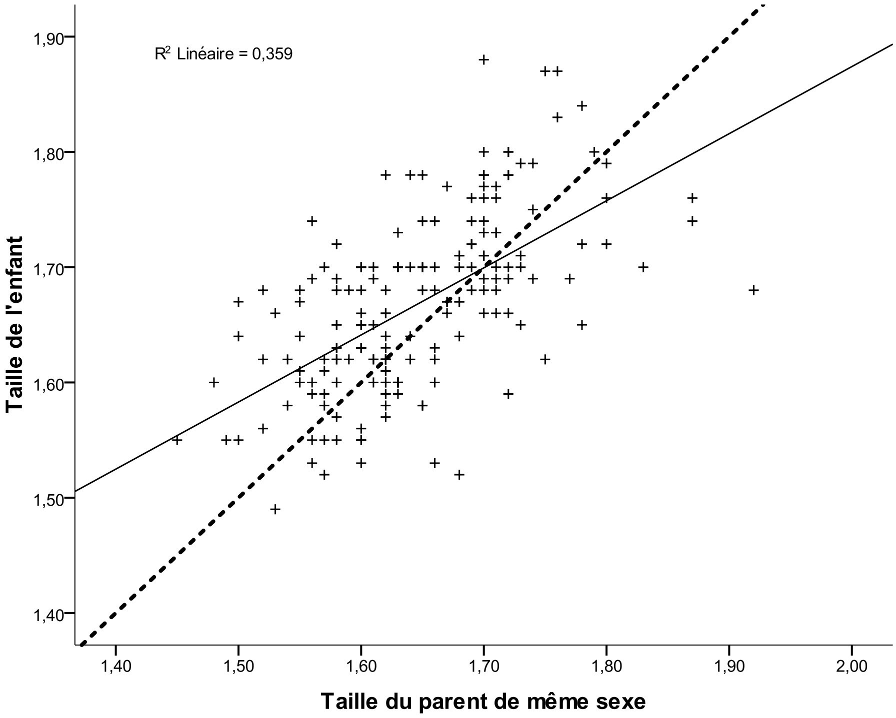


Figure : Mesures réelles sur 182 couples parents enfants

y (taille enfant) = 0.582 X taille parent) + 0.71 avec r = 0.60



°°°°°°°°°°°°°°°°°°°°°°°°°°°°°°°°°°°