Statistiques pour la psychologie — Licence
Régression linéaire

Nicolas Gauvrit
Université de Metz
`http ://adems.free.fr/`

22 décembre 2004

Exercice 1 J. Klatzman[1] *a relevé, pour différents pays, des indices de* consommation d'alcool *par habitant* (A) *et l'*espérance de vie (E) *(base 100 pour le Japon). Les résultats sont les suivants :*

pays	*Nigéria*	*Brésil*	*Chine*	*Japon*	*USA*	*France*
A	2	5	6	9	15	16
E	56	68	71	100	98	95
Pays	*Italie*	*Soudan*	*G.B.*	*Uruguay*	*Lux.*	*Suède*
A	15	1	18	6	14	13
E	94	50	99	60	95	96

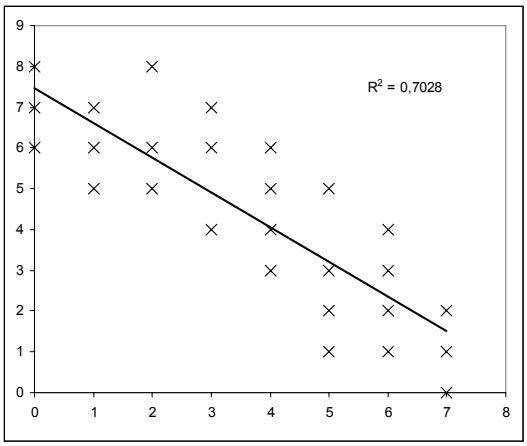
1. *Décrivez la situation statistique et représentez les données.*
2. *Les deux variables A et E sont-elles liées ? Expliquez le lien éventuel.*

□ (1) Les individus sont des pays, et l'on étudie deux variables numériques, à savoir la consommation d'alcool A (variable indépendante) et l'espérance de vie E (variable dépendante). On donnerait le diagramme de dispersion de E en A
(2) Le coefficient de corrélation ($r = .91$) est significativement positif[2], les deux variables sont donc liées positivement. La consommation d'alcool a donc "un effet" positif sur l'espérance de vie. Bien entendu, cet effet n'est pas une relation de cause à effet. En regardant plus attentivement le diagramme de dispersion et la situation des individus (donc des pays) particuliers, on s'aperçoit que les pays pauvres donnent des valeurs faibles pour A et E, alors que les pays riches donnent des valeurs élevées pour ces deux grandeurs. C'est très probablement la *richesse* du pays qui explique à la fois les variations de A et E. □

Exercice 2 *On relève sur tous les élèves d'une école le nombre X de* fautes d'orthographe *à une dictée (la même dictée pour tous) et la* pointure Y *(taille*

[1] dans *Attention ! statistiques,* éd. La découverte, 1997.
[2] Il faudrait en toute rigueur faire un test sur r, mais on peut considérer en première approximation qu'un coefficient $r \geq .5$ est significatif.

des chaussures, unité arbitraire). On trouve les valeurs représenteés ci-dessous :



Concluez (on admettra que le lien entre x et y est négatif). Construisez un diagramme bayesien (causal) rendant compte de la situation.

□ Il y a un lien décroissant entre les deux variables : les élèves ayant de plus grands pieds font donc moins de fautes d'orthographe que les autres. Le coefficient de détermination est même très élevé puisqu'il indique de 70% des variations dans le nombre de fautes d'orthographe s'expliquent par la taille des pieds...
Les données montrent une corrélation qui n'est évidemment pas une relation de cause à effet. En réalité, c'est probablement l'âge — ou le niveau scolaire — qui explique le lien. Les élèves les plus avancés (dans les plus grandes classes) ont des pieds plus grands *et* font moins de fautes d'orthographe.
★ On pourrait représenter les choses par un *diagramme causal* où les flèches indiquent des liens de cause à effet — la flèche allant de la cause vers l'effet — entre les variables :

orthographe
↑
âge
↓
pointure.

□

Exercice 3 *Dans une étude portant sur des jugements de* beauté, *on demande à des volontaires de noter, par une note X, six portraits (A,B,...,F). Deux groupes sont formés. Dans le groupe 1, la consigne est de donner une note comprise en 0 à 10, avec 0 pour "très moche" et 10 pour "très beau". Dans le groupe 2, la consigne est de noter entre -5 (très moche) et 5 (très beau). On notera pour chaque visage représenté, X_1 la note obtenue dans le groupe 1 et X_2 celle obtenue dans le groupe 2.*

1. *Sous certaines hypothèses à préciser, on s'attend à avoir $r_{X_1X_2} = 1$. Expliquez.*
2. *On trouve en réalité $r = 0.91$ et l'équation de régression*

$$\hat{X}_2 = 1.3X_1 - 5.$$

Concluez.

□ (1) Si on suppose que la transposition d'une échelle à l'autre est faite selon une loi simple, à savoir

$$X_2 = X_1 - 5,$$

alors on devrait avoir $r = 1$.
(2) Le coefficient r est très proche de 1. Le lien entre les deux évaluations est

donc linéaire et croissant. L'équation de régression montre cependant que X_2 est, par rapport à X_1, plus "étalée" (puisque le coefficient directeur de la droite est supérieur à 1). On peut donc penser que, par rapport à l'échelle $[-5, 5]$, l'échelle $[0, 10]$ est plus "écrasée", sans doute parce qu'il est psychologiquement plus difficile de noter 0 que -5, même si le "sens" est le même (très moche). On pourra voir là le résultat de l'image négative du 0 — notations scolaires, etc.
(3) Dans cette étude, les visages sont les individus, et les deux variables sont X_1 et X_2. Pour que la régression linéaire de X_2 et X_1 ait un sens, il faut considérer que X_1 et X_2 sont de vraies variables. Il serait très gênant que ces variables dépendent du choix des groupes. Implicitement, on a fait l'hypothèse que X_1 et X_2 étaient de bonnes estimations des variables Y_1 et Y_2 que l'on obtiendrait de la même manière en interrogeant *toute la population*. Autrement dit, on a supposé que les groupes étaient assez grands pour qu'il soit possible de considérer X_1 et X_2 comme des mesures de "la beauté" en général.
★ Les conclusions que nous tirons sur les dispersions devraient être soumises à un examen plus attentif : la régression est en effet une méthode non symétrique, et la régression de X_1 en X_2 ne donne pas, sauf cas particulier, les mêmes résultats que celle de X_2 en X_1... Une méthode *symétrique* qui donnerait l'équation $\hat{X}_2 = 1.3X_1 - 5$ serait bien plus convaincante. □

Exercice 4 *On relève*[3] *chez des élèves l'image de soit par une note X et les résultats scolaires par une note Y. On trouve $r_{XY} = .59$. Concluez. Pourquoi Y est-elle la variable dépendante ?*

□ Le lien éventuel entre les variables X et Y peut s'expliquer de plusieurs manières, et en particulier il peut s'expliquer par une relation causale directe dans les deux sens : un élève qui a de bonnes notes verra probablement son "image de soi" renforcée. Réciproquement, les élèves ayant une bonne image d'eux-mêmes sont réputés avoir de bonnes notes, notamment parce qu'ils ne doutent pas d'eux-mêmes et sont donc moins enclin à perdre du temps à se demander s'ils ont bien répondu. Y a été choisie comme variable dépendante par les chercheurs parce que ces derniers privilégient la seconde approche.
Le coefficient de corrélation montre bien un lien croissant entre les deux variables. Mais il ne permet évidemment pas de dire laquelle des deux explications est juste, si tant est que l'une des deux le soit.
★ La vision choisie par les chercheurs est plus séduisante, mais peut-être aussi moins réaliste, que la vision inverse. Mais il y a probablement un peu des deux...
□

Exercice 5 *Des chercheurs*[4] *ont étudié les résultats en sciences (S) et en* orthographe *(O) d'élèves de primaire, pour les comparer selon le genre*[5] *G*[6]*. Ils trouvent les résultats suivants*

G	S	O	G	S	O
0	3	4	1	2	3
0	4	5	1	5	4
0	5	5	1	6	5
0	6	5	1	7	4
0	7	6	1	8	6
0	7	7	1	8	9
0	6	7	1	9	4
0	8	8	1	8	2

1. Décrivez la situation statistique de deux manières différentes :

[3] Monteil, J.-M. & Hughet, P. (2002). *Réussir ou échouer à l'école : une question de contexte ?* Grenoble : Pug.
[4] Bacharach, V. R. *et al.* (2003). Racial and gender science achievment gaps in secondary education. *The Journal of Genetic Psychology, 144(1)*
[5] Les auteurs étudient aussi les différences blancs/noirs, que nous n'évoquons pas ici.
[6] On notera 0 pour "fille" et 1 pour "garçon".

(a) en supposant qu'il y a une variable "matière" avec pour modalité S et O

(b) en supposant que les élèves sont les individus statistiques

2. Concluez.

□ (1) Une description cohérente mais très peu pratique de la situation statistique est la suivante : les individus sont des copies, et l'on étudie les variables "matière", "note", "sujet" — l'auteur de la copie —, et "sexe". On est donc amené à étudier 4 variables dont une (sujet) est très peu pratique parce qu'elle est nominale avec un grand nombre de modalités.
Mais une description standard et beaucoup plus simple est la suivante : les individus sont des écoliers, et l'on étudie les variables G, S et O. Dans la suite, on considère O comme une variable indépendante et S comme une variable dépendante — de manière totalement arbitraire.
(2) Le premier tableau de résultats montre que les résultats d'orthographe et de sciences sont liés de manière croissante ($r = .87$) chez les filles ($G = 0$). La note d'orthographe explique linéairement 75% des variations de S $\left(r^2 = .75\right)$. La dernière colonne indique que l'on peut s'attendre, en utilisant l'équation de régression pour estimer S à partir de O, à une erreur de 1 point sur la note de science.
Le second tableau montre que les résultats sont similiaires chez les garçons, avec toutefois un lien moins net entre les deux variables, et une erreur d'estimation d'environ 2 points.
Globalement, le troisième tableau montre un lien positif entre les deux notes.
On peut donc résumer tout cela en disant que les notes S et O sont liées positivement chez les élèves aussi bien globalement que chez les filles seules ou chez les garçons seuls, c'est-à-dire que certains élèves sont bons dans les deux disciplines alors que d'autres sont plutôt faibles (par rapport au groupe) dans les deux matières. Cela laisse penser que des compétences communes sont nécessaires à la compréhension des deux domaines au niveau primaire — cela pourrait être, tout simplement, l'acquisition des capacité de lecture et de compréhension de l'écrit.
Il faut cependant noter que le lien est nettement plus fort chez les filles, et qu'il n'est manifestement pas significatif chez les garçons puisque le coefficient de détermination ajusté est négatif $\left(r^2_{aj} = -.046\right)$.
★ La connaissance de r sur les groupes des filles et des garçons ne donnent aucune information sur le r global[7].
★ On pourrait ne prendre en compte que les coefficients r^2_{aj}, puisqu'ils constituent *a priori* de meilleures estimations de ρ^2, coefficient de corrélation sur la population toute entière. □

Exercice 6 *On étudie le lien entre le stress et la santé au moyen d'une mesure numérique X de stress et d'une mesure de l'importance des symptômes Y, également numérique. Sur un échantillon de $n = 107$ personnes, on trouve $r_{xy} = .506$. Calculez le coefficient de corrélation ajusté*

$$r_{aj} = \sqrt{1 - \frac{(1 - r^2)(n - 1)}{n - 2}}$$

et concluez. Cela prouve-t-il que le stress est mauvais pour la santé ?

□ On a trouvé sur notre échantillon un coefficient de corrélation positif, qui indique donc que les symptômes physiques ont tendance à augmenter avec le stress, autrement dit que les personnes plus stressées ont également, généralement, plus de symptômes physiques. Deux remarques doivent être faites ici. D'abord, un lien croissant entre les valeurs X et Y peut autant s'expliquer par un effet du stress sur les symptômes que par un effet inverse des symptômes

[7] *Sauf* que le r global ne peut valoir ni 1 ni -1.

sur le stress[8] et en second lieu, la valeur de r^2 observée est trop optimiste — en moyenne, r^2 calculé sur un échantillon dépasse ρ^2 —, c'est pourquoi il est préférable pour estimer la grandeur de l'effet (ou la force du lien) d'utiliser r^2_{aj}. On trouve ici

$$\begin{aligned} r^2_{\text{aj}} &= 1 - \frac{\left(1 - .51^2\right) \times 106}{105} \\ &= .25. \end{aligned}$$

Le stress "explique" donc 25% des variations de symptômes physiques observés, ce qui est étonnamment élevé, et ne serait peut-être plus vrai dans une population plus hétérogène du point de vue de la santé par exemple — on imagine que les sujets de l'expérience sont tous relativement bien portant. □

Exercice 7 *On demande à 35 volontaires de dessiner un* chapeau *tel que la hauteur h soit égale à la largeur l. Un régression linéaire de l en h donne les résultats suivants : Le coefficient de détermination est .7785 et l'équation de régression est*

$$\hat{l} = .98h + 8.34.$$

Concluez.

□ Le coefficient de détermination est élevé, et le coefficient de corrélation est positif (la pente de la droite, .98, étant positive). Cela est d'un intérêt limité, puisqu'on s'attendait évidemment à trouver $l \simeq h$, et donc une relation linéaire presque parfaite entre l et h (les gens arrivent *a peu près* à dessiner un chapeau carré).
C'est l'équation de régression, bien plus que le coefficient de corrélation, qui nous intéresse ici. L'équation est en effet — à peu près —

$$\hat{l} = h + 8,$$

ce qui montre d'abord que l semble généralement surestimée : lorsque les sujets essaient de dessiner le chapeau carré, il le font en réalité plus large que haut.
Il n'est pas étonnant que les sujets surestiment la largeur : on n'a pas, en général, des résultats parfaits. La forme du chapeau peut induire des biais, de même que son orientation. En revanche, on s'attendrait tout naturellement à avoir une relation entre l et h de la forme $l = \alpha h$ où α est une constante multiplicative. C'est ce qu'on aurait si, par exemple, les sujets avaient tendance à ajouter 5% à la largeur théorique (donc la hauteur). Mais on obtient au contraire une constante *additive,* ce qui indique que les sujets ont tendance, indépendamment de h — entre 20 et 60 mm — à ajouter la longueur constante de 8mm à l !
★ Une alternative à la régression qui traiterait l et h de manière symétrique serait préférable. □

Exercice 8 [9] *Des auteurs anglais mesurent par un grandeur X numérique les capacités de déductions sur les rapports humains (*theory of mind*) chez des sujets d'âge A. On trouve les moyennes conditionnelles suivantes :*

A	*60*	*65*	*70*	*75*	*80*
X	*4*	*4.1*	*3.7*	*3.3*	*2.9*

Une régression linéaire vous paraît-elle judicieuse ? Traitez les données et fournissez une conclusion en supposant que l'on a 10 sujets par âge, et un écart type conditionnel $\sigma_{x|A}$ constamment égal à 1.

□ La courbe des moyennes conditionnelles de X en A montre une relation peut-être forte mais pas linéaire entre les deux variables. Une régression linéaire n'est donc pas adaptée. □

[8] En réalité, cette deuxième explication a été contrôlée et réfutée (dans le cadre de l'étude mais pas en général) par les auteurs.

[9] Taylor, E. A. *et al.* (2003). Does performance on theory of mind decline in old age ? *British Journal of Psychology, 43*

Exercice 9 *On reprend le thème de l'exercice 7, mais en demandant aux sujets de dessiner le chapeau couché. On appelle alors l la dimension verticale et h la dimension horizontale du chapeau. On trouve les résultats suivants :*

$$\begin{aligned} r &= .90 \\ \hat{y} &= .92x + 6.65. \end{aligned}$$

Concluez.

□ On est dans une situation analogue à celle de l'exerice 7, et les résultats sont similaires. Il faut noter deux choses :
C'est encore la largeur du chapeau (ici verticale) qui se trouve surestimée. Ce n'est donc pas en premier chef l'orientation du chapeau qui détermine le biais, mais plutôt sa forme.
La constante additive est plus petite dans cet exemple que dans le cas précédent (exercice 7), ce qui laisse supposer que l'orientation du chapeau intervient également, mais moins.
★ Si on devait *mesurer* les effets, on pourrait par exemple penser que la forme du chapeau entraîne un biais d'environ 7mm (à ajouter à la "largeur" du chapeau), alors que son orientation implique un effet de 1mm (à ajouter à la dimension horizontale). Les effets s'ajoutent dans un cas (exercice 7), donnant au total un biais de $7 + 1 = 8$ millimètres. Au contraire, ils sont opposés dans le second cas (cas présent), donnant un effet global d'environ $7 - 1 = 6$ millimètres. □