Comment bien régresser

La statistique peut-elle se passer d'artefacts ?[1]

Jean-Bernard Chatelain

Prisme N° 19

Octobre 2010

[1] Ce texte est la transcription du séminaire « Probabilismes » de mai 2010, dont Xavier Ragot était le discutant. Qu'il soit remercié pour l'ensemble de ses critiques et de ses encouragements. Ce texte présente des idées inspirées d'un travail commun avec Kirsten Ralf.

.

Résumé

Cet article présente un cas particulier de régression non fondée, lorsqu'une variable dépendante a un coefficient de corrélation simple proche de zéro avec deux autres variables, qui sont en revanche très corrélées entre elles. Dans ce type de régressions, les paramètres mesurant la taille des effets sur la variable dépendante sont très élevés. Ils peuvent être « statistiquement significatifs ». Publier en priorité des résultats dans une revue scientifique qui ont cette propriété est l'une des raisons pour lesquelles les régressions fausses sont nombreuses, d'autant qu'il est facile de les construire avec des variables retardées, mises au carré, ou interagissant avec une autre variable. De telles régressions peuvent contribuer à la renommée des chercheurs, en stimulant l'apparition d'effets élevés entre variables. Les effets, souvent inattendus, sont fragiles. Ils dépendent souvent de quelques observations, ce qui donne l'occasion de lancer des controverses scientifiques. Ces controverses empiriques se soldent par la confirmation de l'absence d'un effet lors des méta-analyses faisant la synthèse statistique de la littérature évaluant cet effet entre deux variables. Nous donnons un exemple de ce phénomène dans la littérature empirique visant à évaluer l'effet de l'aide au développement sur la croissance économique.

Sommaire

Introduction

L'une des méthodes statistiques les plus utilisées dans les sciences appliquées est la méthode de régression linéaire. Elle permet d'évaluer que la hausse d'une variable (par exemple, l'aide au développement) est associée à un effet plus ou moins important à la hausse ou à la baisse sur une autre variable (par exemple, la croissance économique). Son origine remonte à la méthode des moindres carrés des erreurs (Legendre, 1805). Elle est ensuite associée à la corrélation entre deux lois normales par Galton (1886) dans le cas d'une régression dite simple. L'extension aux corrélations partielles d'une variable dépendante avec au moins deux autres variables est inventée par Yule (1897). Elle est appelée régression multiple.

Ce texte remet en cause la validité de certains résultats obtenus par la méthode de régression linéaire. Dans un premier temps, nous ajoutons à une liste déjà longue (Aldrich, 1995) un nouveau cas de régression non fondée, où la méthode de régression linéaire indique une liaison entre plusieurs variables, alors que cette liaison n'est pas vérifiable. Ce type de régression se présente comme un cas très particulier, voire anecdotique, que les chercheurs ne devraient rencontrer que rarement dans leur pratique. Nous argumentons que le mode de sélection des publications utilisant la régression linéaire présente un artefact favorisant la publication de ces régressions non fondées.

Pour ce faire, il faut d'abord évoquer la question de l'inférence statistique : comment peut-on tirer, à partir d'un certain nombre d'observations, une certaine probabilité sur des relations qui peuvent exister entre plusieurs variables. Une méthode d'inférence adaptée à la méthode de régression a été proposée par Fisher (1925). Elle s'est progressivement imposée auprès des chercheurs. En dépit de son extrême diffusion dans la plupart des communautés scientifiques, elle fait l'objet de nombreuses critiques qui sont brièvement rappelées.

Dans un troisième temps, nous décrivons ce que la méthode d'inférence statistique proposée par Fisher a engendré comme pratique chez les chercheurs qui utilisent cette méthode. Elles ont conduit les éditeurs de revues scientifiques à ne publier que les résultats pour lesquelles l'inférence statistique conduit à rejeter l'hypothèse dite « nulle » d'absence de liaison entre deux variables. Ce problème conduit à un artefact de publication où les résultats « négatifs », indiquant l'absence de relation entre variable ne sont pas publié.

Le cas particulier de régression non fondée mentionné ici devient alors très intéressant. Il permet de trouver des paramètres mesurant la liaison statistique entre deux variables (1) très élevés, (2) rejetant l'hypothèse nulle d'absence de liaison entre les variables, (3) dont la valeur estimée et le signe sont particulièrement sensibles à l'ajout ou au retrait de quelques observations, (ce qui stimulera des controverses et la notoriété du chercheur) et (4) révélant une liaison *a priori* imprévue entre ces variables, ce qui est valorisé dans les revues scientifiques les plus prestigieuses.

Dans la quatrième section, nous proposons deux remèdes simples : le facteur d'inflation d'un paramètre, destiné à mesurer la taille de l'effet entre deux variables, et des tests d'hypothèses sur les coefficients de corrélations simples. Nous utilisons ces indicateurs sur une étude sur les effets de l'aide au développement sur la croissance économique extrêmement citée ces dix dernières années. À l'aide de ces indicateurs, nous vérifions qu'il s'agit d'une régression non fondée.

1. Le problème négligé de régressions non fondées.

a. Un artefact de la régression multiple

La régression simple estime une relation linéaire entre deux variables observées *N* fois : par exemple, la croissance du produit intérieur brut (PIB) par habitant et l'aide au développement rapportée au PIB observés pour *N* pays durant une certaine période.

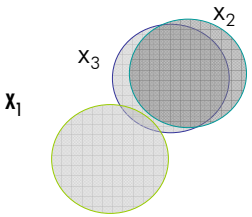
Pour présenter le problème, nous considérons des variables « standardisées », de moyenne nulle et d'écart-type égal à l'unité. L'écart-type est une mesure de la dispersion des observations autour de la moyenne. On peut toujours standardiser des variables en soustrayant aux observations leur moyenne et en divisant le tout par leur écart-type. Pour des variables standardisées, la méthode de régression simple estime une relation linéaire entre deux variables, dont le paramètre est le coefficient de corrélation simple, tel que défini par Galton (1886). Plus ce coefficient de corrélation est élevé en valeur absolue, plus la « taille de l'effet » d'une variable sur l'autre est forte. Sa valeur absolue est au plus égale à l'unité. Voici trois exemples de régression simple :

$$x_2 = 0{,}99\, x_3 + \varepsilon_{2.3}$$
$$x_1 = 0\, x_2 + \varepsilon_{1.2}$$
$$x_1 = 0{,}14\, x_3 + \varepsilon_{1.3}$$

La variable à gauche de l'équation est appelée « variable dépendante » ou variable expliquée. La variable à droite de l'équation est appelée « variable explicative ». Des erreurs de mesures et de variables omises sont prises en compte dans les « perturbations » notées $\varepsilon_{i.j}$. Leurs valeurs numériques pour chaque observation de l'échantillon sont appelées « résidus ».

On peut interpréter le coefficient de corrélation comme suit. Lorsque la variable x_3 s'écarte de sa moyenne d'un écart-type, alors la variable x_2 s'écarte de sa propre moyenne de 0,99 fois son écart-type. Les deux variables sont très corrélées. En revanche, pour la deuxième équation, lorsque la variable x_2 s'écarte d'un écart-type de sa moyenne, alors la variable x_1 ne s'écarte pas de sa moyenne. Les deux variables ne sont pas du tout corrélées. Enfin, les variables x_1 et x_3 sont très faiblement corrélées. Ces corrélations peuvent être représentées par un diagramme de Venn (Figure 1.1), où le disque représentant x_2 couvre en grande partie le disque représentant x_3 sans avoir d'intersection avec le disque représentant x_1, qui a une petite intersection avec le disque x_3.

Figure 1.1 : Diagramme de Venn pour les trois corrélations simples.



L'analyse de la variance de la variable dépendante complète ces informations. La variance est le carré de l'écart-type, lui-même noté σ. Dans le cas d'une variable standardisée, la variance est égale à l'unité. L'analyse de la variance indique la répartition de la dispersion des observations suivant qu'elle est prédite par la dispersion de la variable explicative (variance expliquée par le modèle) ou par la perturbation associée, par exemple à des erreurs de mesures où à d'autres phénomènes non observés (variance résiduelle). On définit également le coefficient

de détermination comme le ratio de la variance de x_2 qui est « expliquée » par la variable x_3 divisée par la variance de la variable dépendante. Dans la régression simple, le coefficient de détermination est le carré du coefficient de corrélation.

Dans le cas de variables standardisés ($\sigma^2(x_1) = \sigma^2(x_2) = \sigma^2(x_3) = 1$), les calculs sont très simples :

$\sigma^2(x_2) = 0{,}99^2\, \sigma^2(x_3) + \sigma^2(\varepsilon_{2.3})$, $R^2 = 0{,}99^2 = 98\%$ $\sigma^2(\varepsilon_{2.3}) = 1 - R^2 = 0{,}02$

$\sigma^2(x_1) = 0^2\, \sigma^2(x_2) + \sigma^2(\varepsilon_{1.2})$ $R^2 = 0\%$ $\sigma^2(\varepsilon_{1.2}) = 1$

$\sigma^2(x_1) = 0{,}14^2\, \sigma^2(x_3) + \sigma^2(\varepsilon_{1.3})$ $R^2 = 2\%$ $\sigma^2(\varepsilon_{1.2}) = 0{,}98$

Sur la base de ces trois coefficients de corrélation simple et des coefficients de détermination, Galton aurait probablement déduite en 1886, qu'il n'existe pas de lien entre les variables x_1 et les deux variables x_2 et x_3, ou que ce lien est négligeable. Yule (1897) étend la méthode de régression linéaire au cas de plusieurs variables (on parle alors de régression multiple). Pour les trois coefficients de corrélations simples de l'exemple ci-dessus, on obtient les résultats suivants en utilisant les formules de Yule (1897). Il note r_{12} le coefficient de corrélation entre la variable x_1 et la variable x_2.

$x_1 = -7{,}01\, x_2 + 7{,}08\, x_3$

$R^2 = -7{,}01 \times r_{12} + 7{,}08 \times r_{13} = -7{,}01 \cdot 0 + 7{,}08 \times 0{,}14 = 100\%$

Quelle surprise ! Les deux variables qui n'avaient aucun effet ou un effet négligeable dans les régressions simples sur la variable x_1 expliquent désormais 100% de la variance de x_1 lorsqu'elles interviennent simultanément dans la régression multiple. Les coefficients sont très élevés pour des variables standardisées. Lorsque la variable x_2 s'écarte d'un écart-type de sa moyenne, alors la variable x_1 s'écarte de sa moyenne de -7,01 fois son écart-type, en supposant que la variable x_3 est inchangée (suivant l'hypothèse « toutes autres choses égales par ailleurs » ou *ceteris paribus*). Lorsque la variable x_3 s'écarte d'un écart-type de sa moyenne, alors la variable x_1 s'écarte de sa moyenne de 7,08 fois son écart-type, en supposant que la variable x_2 est inchangée. Il s'agit de réactions extrêmes de la variable dépendante.

On peut calculer les facteurs d'inflation des paramètres (en anglais, « Parameter Inflation Factor » abrégé en *PIF*). Le PIF est un indicateur proposé par Chatelain et Ralf (2010). Il est défini comme le ratio du paramètre estimé obtenu dans une régression multiple divisé par le paramètre estimé par la régression simple. On peut le calculer pour chacune des variables explicatives de la variable x_1 :

$PIF_{1,2} = -7{,}01/0 \qquad PIF_{1,3} = 7{,}08/0{,}14 = 50.$

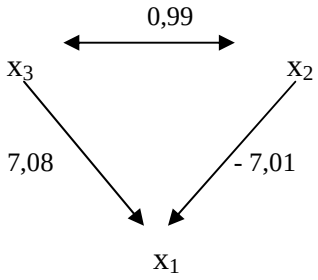
Pour la variable x_2, le *PIF* est infini, tandis que pour la variable x_3, le *PIF* est de 50. Les tailles des effets des deux variables x_2 et x_3 sur la variable x_1 ont été considérablement amplifiées.

Revenons sur l'interprétation, toutes choses égales par ailleurs, des coefficients de la régression multiple. Elle a été initialement proposée par l'économiste américain Moore (1917), par ailleurs, un très grand admirateur de la *personnalité* de Cournot (Moore (1905)). Elle s'est imposée depuis, mais l'exemple ci-dessus indique que son usage systématique n'a pas toujours de sens. En effet, x_2 et x_3 sont très corrélées. Si x_2 s'écarte d'un écart-type de sa moyenne, x_3 s'écartera elle aussi de quasiment un écart-type de sa moyenne d'après la première équation. Pearl (2009, pp. 356–57) suggère qu'il est toujours possible de faire une expérience mentale « *comme si* » x_3 ne bougeait pas, par une *intervention* de x_2, indépendante de la liaison statistique très forte existant entre x_2 et x_3. Encore faut-il que les résultats de cette expérience aient un sens et permettent de justifier des *PIF*s extravagants, rendant compte d'amplifications colossales des paramètres dans des régressions multiples.

b. Une interprétation par l'analyse des chemins causaux

Pour mieux comprendre le problème, faisons l'analyse des chemins des liaisons statistiques « partielles » proposé par le généticien et statisticien Wright (1920) (Figure 1.2).

Figure 1.2 : Chemins causaux avec médiation par x_3.



Il y a un effet partiel direct de x_2 sur x_1, donné par le coefficient -7,01. Il y a aussi un effet indirect partiel de x_2 sur x_1 par l'intermédiaire de x_3. On obtient cet effet comme le produit de l'effet de x_2 sur x_3 que multiplie l'effet de x_3 sur x_1, soit 0,99 x 7,08. L'effet total de x_1 sur x_2 est la somme de l'effet direct et de l'effet indirect : il est égal au coefficient de corrélation simple. Dans le cas présent, l'effet indirect est exactement compensé par l'effet direct, si bien que l'effet total est nul :

$-7{,}01 + 0{,}99 \times 7{,}08 = 0$

Deux interprétations possibles de ce résultat coexistent :

- la régression multiple est fausse : les deux variables x_2 et x_3 mesurent à peu près la même chose, ou sont très liées l'une à l'autre. En réalité, elles n'ont aucune liaison ou un effet négligeable avec la variable x_1. La taille élevée des paramètres de la régression multiple est obtenue principalement du fait de la corrélation élevée entre les deux variables explicatives. Le résultat obtenu est un artefact des formules de la régression multiple obtenu par Yule en 1897, lorsque les variables explicatives sont très corrélées.
- La régression multiple décrit un modèle parfaitement homéostatique, pour suivre une idée importante en médecine mise en avant par Bernard ([1865] 2008) puis reprise par Wiener (1948) dans son exposé de la cybernétique. Une des deux variables, par exemple, la variable x_3, est une variable associée à une rétroaction négative suite à un choc sur l'autre variable explicative x_2. Elle permet d'assurer une stabilité parfaite de la variable x_1 en dépit du choc sur la première variable. Par exemple, la variable x_3 peut être une variable de politique monétaire ou budgétaire visant à diminuer les fluctuations du PIB (Hoover, 2001, pp. 45–6).

Pour Hoover (2001, pp. 45–6), la méthode de régression et les observations des variables ne permettent pas d'établir une distinction entre ces deux interprétations « ontologiquement » différentes. Il faut des informations supplémentaires sur la nature des variables pour choisir quelle interprétation donner aux résultats. Les régressions infondées que nous évoquons ici ne correspondent à aucune des diverses régressions dénoncées depuis Pearson (1897) telles qu'exposées dans Aldrich (1995). Par exemple, Simon (1954) insiste sur des régressions non fondées associées à une corrélation partielle nulle (associées à un paramètre nul

dans une régression multiple) alors que le coefficient de corrélation simple (associé à une régression simple) est non nul.

Notre cas de régression correspond à une discordance en sens inverse de celle de Simon (1954). En effet, le coefficient de corrélation partiel est non nul (il est même très élevé) alors que le coefficient de corrélation simple est nul. Cette situation est identique à la violation de la « *condition de fidélité des graphes causaux* » par Spirtes, *et al.* (2000) aussi appelée « *condition de stabilité des indépendances conditionnelles* » par Pearl (2009, p. 48). Ces auteurs ont poursuivi l'approche initiale de Wright (1920) d'analyse des chemins des liaisons entre variables. Ils proposent des algorithmes permettant de retenir ou d'éliminer des liens causaux potentiels entre des variables. Ces algorithmes sont fondés sur des conditions à satisfaire ou non par les coefficients de corrélations simples et par les coefficients de corrélation partiels.

Par exemple, l'élimination des variables explicatives à corrélation simple nulle avec la variable dépendante est le point de départ de la sélection des variables explicatives dans une version récente de l'algorithme de Spirtes, *et al.* (2000) proposée par Bühlmann, *et al.* (2010). Nous détaillons ce point dans la section 4. Leur article présente une application de leur algorithme à un échantillon où le nombre d'observations est de 71 bactéries produisant de la riboflavine pour 4088 variables explicatives correspondant à autant de niveaux d'expressions de gènes de ces bactéries.

Au début, l'argumentation mise en avant par Spirtes, *et al.* (2000) et par Pearl (2009) est que les violations de la « *condition de stabilité des indépendances conditionnelles* » sont très rares. Plus précisément, une égalité stricte entre paramètres serait de mesure nulle (telle que : $-7{,}01 + 0{,}99 \times 7{,}08 = 0$ sur notre exemple) dans l'ensemble de la distribution des paramètres, lorsqu'ils sont libres de varier indépendamment les uns des autres (Pearl, 2009, p. 62). Autrement dit, les occurrences de l'artefact de la régression multiple dont nous parlons ici devraient être rares.

En revanche, Freedman (1997) avance qu'il n'y a pas de raison de rejeter l'hypothèse que des paramètres soient liés par des contraintes d'égalité. En particulier, Hoover (2001, pp. 45–6) maintient que ce phénomène est très fréquent en économie, du fait des possibilités de contrôle par la politique économique. Il met en avant un diagramme théorique du modèle IS-LM où la variable de politique

monétaire peut rendre inchangée l'activité économique suite à un choc provenant d'une autre grandeur économique. La question posée n'est pas la possibilité théorique du modèle homéostatique. Il s'agit de l'observation de corrélations empiriques où les taux directeurs permettent d'annuler complètement et rapidement les chocs externes sur le PIB, au point de faire disparaître sa variance. À ce titre, l'exemple est mal choisi. Les capacités de réactivité et de stabilisation des banques centrales sur l'activité économiques sont loin d'annuler la variabilité du PIB ou de l'inflation en les rendant exactement non corrélés aux chocs externes. En revanche, Wiener (1948) faisait plutôt part de son étonnement face à des systèmes sociaux et politiques qui ne réussissaient pas à s'autoréguler.

L'argument de cet article est que les occurrences de ces régressions non fondées sont beaucoup plus élevées que ne le présument Spirtes *et al.* (2000) et Pearl (2009). Ceci vient d'une toute autre raison que la possibilité de contrôle parfait dans les modèles homéostatiques mise en avant par Hoover (2000). Comme exposé dans la suite de l'article, cette fréquence provient (1) de la déconnection de l'origine de ces régressions non fondées avec le test d'inférence sur l'existence d'un effet, proposé par Fisher, (2) des critères déterminant le succès des chercheurs. Nombre de chercheurs publient ces régressions non fondées sans s'en rendre compte.

c. Une interprétation par l'orthogonalisation des variables explicatives

Pour aborder le point (1) ci-dessus, nous donnons un autre éclairage concernant ces régressions non fondées. Puisque ce problème est associé à une corrélation trop élevée entre des variables explicatives, on peut le contourner en transformant des variables corrélées en variables non corrélées. Par analogie avec la géométrie euclidienne, cette transformation est parfois appelée l'« orthogonalisation » des variables explicatives. On inclut dans la régression multiple les résidus de la régression entre les deux variables explicatives et on obtient ce résultat :

$$x_1 = 0\, x_2 + 7{,}08\, (x_3 - 0.99\, x_2) \qquad R^2 = 100\%$$

La nouvelle régression multiple obtenue correspond à la mise en facteur du paramètre 7,08 de la première régression multiple. À la différence de la première

régression multiple, il apparaît clairement que x_2 n'a pas d'effet sur x_1. Il ne reste qu'une variable explicative : cette régression est équivalente à une régression simple. Le coefficient de la variable qui reste (la variable x_3 nette de sa corrélation avec x_2) est très élevé (Figure 1.3) :

Figure 1.3 : Chemins causaux après orthogonalisation.

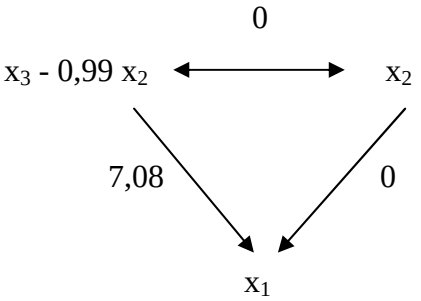
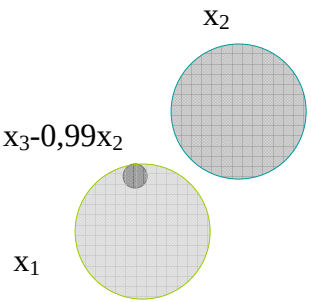


Figure 1.4 : Diagramme de Venn après orthogonalisation.



Le coefficient de détermination R^2 est inchangé : il vaut 1 ainsi que le coefficient de corrélation simple entre la variable x_1 et la variable $x_3 - 0,99\ x_2$. Donc, cette variable « explique » l'intégralité de la variance de x_1 La dispersion des observations de la variable explicative ($x_3 - 0,99\ x_2$) autour de sa moyenne est minuscule. Son écart-type ne vaut plus 1, mais 0,02, soit 2% de l'écart-type de la variable expliquée. Sur le diagramme de Venn (Figure 1.4), le disque représentant la

variance de cette variable est relativement petit. Il est entièrement inclus dans le disque associé à la variable dépendante x_1 car, dans cet exemple extrême, le coefficient de détermination R^2 est égal à l'unité. Enfin le disque représentant la variance de la variable x_2 n'a pas d'intersection avec le disque représentant la variance de la variable dépendante x_1, parce que la corrélation est nulle entre les deux variables, comme dans le diagramme de Venn avant orthogonalisation (Figure 1.1).

Le paramètre de la régression simple 7,08 est désormais un paramètre « non standardisé ». De manière générale, la relation entre un paramètre standardisé (indicé par *S*) un paramètre non standardisé est la suivante :

$$\beta_{12} = \beta_{S,12} \frac{\sigma(x_1)}{\sigma(x_2)}$$

Dans le cas présent, le paramètre non standardisé 7,08 correspond à un paramètre standardisé égal au coefficient de corrélation (il s'agit d'une régression simple) que multiplie le rapport des écarts-types entre la variable dépendante (égal à 1) et l'écart-type de variable explicative.

$$7{,}08 = 1 \times \frac{\sigma(x_1)}{\sigma(x_3 - 0{,}99x_2)} = \frac{1}{\sqrt{1-0{,}99^2}} = \frac{\mathrm{cov}(x_1, x_3 - 0{,}99x_2)}{\sigma^2(x_3 - 0{,}99x_2)}$$

Lorsque la dispersion des observations de la variable explicative autour de sa moyenne est relativement minuscule par rapport à la dispersion de la variable dépendante, la *taille* et le *signe* du paramètre estimé est très instables, suivant qu'on ajoute ou qu'on supprime une observation atypique éloignée de la moyenne de la variable explicative. Cette observation peut faire levier sur la valeur du paramètre de la régression à la hausse ou à la baisse. Dans ce contexte, il est très fréquent que l'ajout de quelques observations modifie fortement le paramètre obtenu, même si le R^2 est très élevé dans l'échantillon initial.

L'interprétation du problème s'est déplacée à l'occasion de l'orthogonalisation. Il ne s'agit plus d'un problème de corrélation entre deux variables explicatives. Il s'agit désormais d'un paramètre estimé très élevé et très sensible aux observations à fort levier.

Ce résultat est loin d'être anodin. L'introduction initiale de deux variables explicatives très corrélées entre elles, et dont la « différence » repose sur quelques observations, permet de transformer un cas particulier en un cas général. Par

exemple, *« Le Botswana est un pays d'Afrique à forte croissance économique à la différence des autres pays d'Afrique. Il a de plus reçu de l'aide au développement »* devient *« L'aide au développement a un effet sur la croissance seulement pour l'ensemble des pays en développement qui ont de « bonnes » politiques macro-économiques »*. La deuxième assertion est beaucoup plus facile à publier (Chatelain et Ralf, 2010).

En conséquence, pour ces régressions non fondées :

(1) l'orthogonalisation des variables explicatives met en évidence l'absence de liaison entre une des variables explicatives et la variable dépendante.

(2) l'explication restante est associée à une variable résiduelle (les résidus d'une régression entre deux variables très corrélées) dont les observations sont très concentrées autour de sa moyenne. En conséquence, le paramètre sera très élevé, mais aussi très sensible à la présence de quelques observations à fort levier.

Par ailleurs, le fait que la dispersion des observations de la variable explicative soit petite est a priori *indépendant* du nombre d'observations. Ce problème d'instabilité des paramètres élevés *ne doit pas être confondu* avec la question de l'inférence statistique abordée dans la section suivante qui prendra en compte le nombre d'observations. En réalité, l'inférence statistique ne sera d'aucun secours face à ce problème. Au contraire, l'utilisation d'un échantillon de grande taille pourra donner à tort l'illusion au praticien que les paramètres élevés sont des résultats solides.

2. L'inférence de l'existence d'un lien entre deux variables

a. De l'induction à l'inférence

L'induction est une méthode décrite, notamment par Aristote, proposant d'établir des relations générales et des prédictions à partir d'un nombre limité d'observations (Milton, 1987). Elle est rejetée par les philosophes sceptiques, tels que Sextus Empiricus ([v.200] 1997) et Hume ([1739] 2000).

Une réponse de statisticiens et de probabilistes au problème de l'induction est d'établir des inférences en associant des probabilités aux fréquences des événements observés (Keuzenkamp, 2000). Une fois calculés les paramètres reliant des variables dans une régression, il faut faire une inférence sur l'hypothèse d'existence d'un lien entre les variables. Ceci revient à tester l'hypothèse de nullité du paramètre associé aux deux variables. L'intuition fréquentiste est qu'un plus grand nombre d'observations pourrait réduire les diverses probabilités de se tromper en réalisant le test.

Fisher (1925) a adapté un test statistique initialement proposé par Student (1908) au test de l'hypothèse de la nullité d'un paramètre dans une régression simple ou multiple. Chaque paramètre estimé de la régression est associé à un écart-type estimé mesurant l'incertitude concernant la valeur du paramètre qui est une fonction décroissante du nombre d'observation noté *N*. Pour la régression simple, la statistique de Student rapporte le paramètre estimé à son écart-type estimé, et si cette statistique dépasse un seuil fixé à l'avance, noté $t_{N,1-\frac{\alpha}{2}}$, en pratique, de l'ordre de 1,96 lorsque l'échantillon dépasse 100 observations, on considère qu'on a moins d'une chance sur vingt ($\alpha = 5\%$) de se tromper en rejetant l'hypothèse de nullité du paramètre conditionnellement à ce que, pour le vrai modèle, cette hypothèse de nullité du paramètre soit vraie (ce qu'on appelle la probabilité de l'erreur de type I, avec la notation $p<0{,}05$).

(2.1) $$t_{12} = \frac{r_{12}}{\sqrt{1-r_{12}^2}}\sqrt{N-2} > t_{N,1-\frac{\alpha}{2}}$$

Par rapport à la section précédente, une nouvelle donnée intervient : le nombre d'observations. Dans la Figure 2.1, l'axe vertical est le nombre d'observations et l'axe horizontal est le coefficient de corrélation. La zone critique correspondant au rejet de l'hypothèse de nullité du coefficient est au-dessus dessus d'une courbe en forme d'entonnoir, donnée par le cas d'égalité dans l'équation (2.1). On ne rejette pas l'hypothèse de nullité du paramètre à l'intérieur de l'entonnoir.

Figure 2.1 : Zone critique du test de Student d'une régression simple.

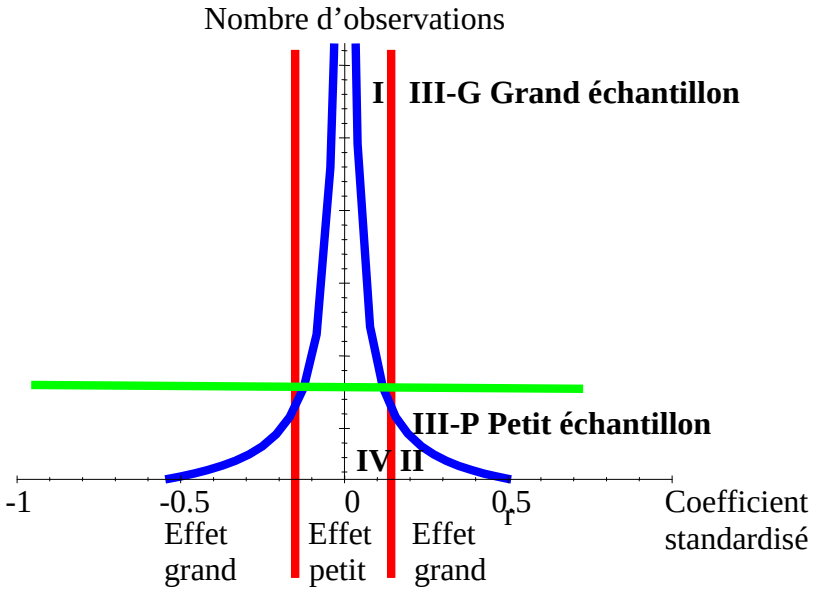


Tableau 2.1 Discordances entre significativité substantive et significativité statistique.

	Petit échantillon (et petite population totale)	Grand échantillon
Petite taille de l'effet (effet « négligeable »)	Zone IV : existence de l'effet rejetée et effet de taille négligeable	**Zone I : existence de l'effet non rejetée, alors que sa taille est négligeable**
Grande taille de l'effet	**Zone II : existence de l'effet rejetée alors que sa taille est importante.**	Zone III : existence de l'effet non rejetée et taille de l'effet non négligeable.

L'inférence à la Fisher s'est imposée systématiquement pour les publications scientifiques dans de nombreux domaines des sciences appliquées : économie, épidémiologie, écologie, marketing, sciences de l'éducation, etc. À la manière du titre ironique d'un article du statisticien J. Cohen (1994) (« *La terre est ronde :* $p<0.05$ »), toute hypothèse scientifique devrait vérifier que la probabilité de l'erreur de type I est inférieure à 5% pour être considérée comme valide.

Un tel tour de force méthodologique sur la validation universelle de vérités scientifiques établissant des liens entre des variables ne peut être que controversé. De manière générale, peut-il exister un critère unique de vérité pour l'induction ou pour l'inférence statistique ? Le contre-argument de Sextus Empiricus ([v.200] 1997, II.4.19), repris par Hume ([1739] 2000), est difficilement contournable : « *Pour que le désaccord qui existe sur le critère fasse l'objet d'une décision, il faut que nous ayons un critère sur lequel nous soyons d'accord* », et ainsi de suite, ce qui conduit soit à une régression à l'infini dans la recherche d'un nouveau critère validant le critère de l'étape précédente, soit à reprendre le premier critère pour qu'il se valide lui-même, ce qui conduit à un argument circulaire.

En conséquence, le débat entre statisticiens sur d'autres critères que celui de Fisher a eu lieu dès ses origines. Student (alias Gosset), E. Pearson et J. Neyman, défendent un avis différent sur l'inférence à la Fisher (McCloskey et Ziliak, 2008). Nous rappelons brièvement quelques-unes des critiques les plus importantes pour les praticiens.

Les chercheurs souhaitent en général obtenir des résultats à l'extérieur de l'entonnoir : c'est là qu'ils peuvent faire l'inférence que le paramètre est non nul selon Fisher. Dans ce cas, le paramètre est appelé, pour reprendre le vocabulaire de Fisher, *statistiquement significatif* ou significativement différent de zéro. Fisher a donc utilisé le terme « significatif », mais dans le langage courant, aussi bien en français qu'en anglais, un effet significatif est un effet qui est fort. Le critère du langage courant, aussi appelé significativité « substantive » (« qui a du sens »), correspond à un coefficient de corrélation élevé, indépendamment du nombre d'observations de l'échantillon, par exemple au dessus (en valeur absolue) des valeurs indiquées par les deux droites verticales rouges sur la Figure 2.1. La significativité statistique de Fisher ne correspond pas aux mêmes zones que la

significativité substantive. On a donc deux types de discordances entre ces notions de significativité (Tableau 2.1).

Dans l'aire II du graphique 2.1 se trouve les cas où l'effet est important mais où la population totale et homogène est toute petite. Par conséquent, *un échantillon homogène* ne peut qu'être petit. On peut penser au cas d'un traitement pour des maladies orphelines en médecine, ou au cas où une des variables correspond à des caractéristiques institutionnelles qui n'apparaissent que dans une vingtaine de pays développés. Même si l'effet (le coefficient de corrélation) est très fort, si l'échantillon est trop petit, on n'arrivera jamais à obtenir une inférence selon laquelle ce paramètre est statistiquement différent de zéro ($p<0.05$).

Inversement, dans l'aire I du graphique 2.1., si l'échantillon est très grand, on pourra considérer qu'un effet minuscule ou négligeable de la variable x_2 sur x_1 est « statistiquement significatif » ($p<0.05$). Dire que cet effet est tout petit, cela signifie qu'un choc d'un écart-type de x_2 par rapport à sa moyenne, fait très peu dévier la variable x_1 de sa moyenne. Dans ce cas, il suffit d'étendre les échantillons indéfiniment, *en agrégeant au passage des populations très hétérogènes,* pour obtenir la significativité statistique d'effets minuscules.

b. Sortir du conflit entre significativité substantive et significativité statistique

Une première réponse à ce débat est de dire que l'hypothèse simple d'existence d'un effet, ($r = 0$) à tester n'est pas celle qui a plus de sens. Il conviendrait de tester une hypothèse composite $r > r(seuil\ minimal)$ en précisant un seuil minimal en dessous duquel l'effet est considéré comme négligeable. Se pose alors la question de la détermination de ce seuil minimal. Dans certains domaines, on pourrait considérer qu'un seuil minimal pour un coefficient de corrélation d'au moins 0,1 (soit un coefficient de détermination expliquant au moins 1% de la variance dans une régression simple. Dans certaines activités, financières par exemple (marché des changes, marché obligataire), de toutes petites variations des cours peuvent conduire à des gains considérables. En physique, de toutes petites variations de certains phénomènes peuvent avoir des conséquences considérables pour valider des théories. La notion de ce qui est un effet négligeable dépend du contexte. Comment faire ?

De plus, l'approche de Fisher a elle aussi un arbitraire concernant un autre seuil, celui de la probabilité de l'erreur de type I (la p-valeur) : pourquoi retenir 5% ? Elle néglige également une autre p-valeur, la probabilité d'une erreur de type II. L'erreur de type II peut être calculée lorsqu'on connaît les hypothèses différentes de l'hypothèse initiale testée (par exemple, toutes les valeurs possibles des coefficients de corrélation non égaux à zero). De surcroît, lorsque l'on fait varier le seuil sur le p-valeur de l'erreur de type I, la probabilité de l'erreur de type II se modifie.

Student avait proposé un autre critère à minimiser pour décider du seuil des deux types de p-valeur, et aussi, potentiellement, du seuil minimal pour un paramètre. Il introduit une *fonction de perte* faisant intervenir deux coûts spécifiques à chacun des deux types d'erreur (McCloskey et Ziliak, 2008). L'introduction de ces coûts spécifique prend en compte le « contexte » de la décision. La fonction de perte la plus simple est l'espérance du coût total des deux types d'erreur. Dans ce cas, on calcule la moyenne des coûts pondérée par les probabilités de chacun des deux types d'erreur. Cette pratique est par exemple proche de celles des banques octroyant ou non un crédit. Le coût de ne pas donner un crédit à quelqu'un qui va rembourser (erreur de type II) est la perte d'une marge d'intermédiation que multiplie la taille du crédit. Le coût de donner un crédit à quelqu'un qui ne va pas rembourser (erreur de type I) est la perte du montant du crédit et des intérêts. Le second coût est plus élevé que le premier. La banque choisira une probabilité de l'erreur de type I beaucoup plus faible que la probabilité de l'erreur de type II : elle sera plus restrictive sur ses crédits.

Les critiques de l'approche de Fisher considèrent qu'elle ne minimise pas une fonction de perte pour les disciplines scientifiques, en particulier pour ses conséquences pratiques pour les décisions à prendre en économie, en médecine, etc. Les décisions et les actions qui s'ensuivront en utilisant des résultats fondés sur le critère de Fisher ne seront pas optimales, puisque, par construction (« $p < 0.05$ »), celui-ci ne prennent pas en compte les coûts relatifs des deux types d'erreurs.

Cette discussion entre le critère de Fisher et celui d'un fonction de perte associée aux conséquences d'une décision, comme proposée par Student, est analogue à une distinction entre deux critères par Sextus Empiricus ([v. 200] 1997, 1.11.21): « *On parle de critère en deux sens : celui que nous prenons pour nous*

convaincre de l'existence ou de la non-existence de quelque chose (…), et celui qui concerne l'action : en nous y attachant, nous feront telles chose, et nous ne ferons pas telles autres. ». Le premier critère est rejeté par le philosophe sceptique. En revanche, il accepte d'utiliser le second critère, afin de ne pas être condamné à l'inaction.

Quel rapport y a-t-il entre la régression fausse de la section 1 et les critiques de l'approche de Fisher de la section 2 ? Elle ne correspond ni à la zone I ni à la zone II. Les effets obtenus par la régression multiple sont très *forts* (lorsqu'on les interprète toute chose égale par ailleurs) et ils peuvent être *statistiquement significatifs*, alors qu'il n'existe pas de liaison entre x_1 et x_2. Ces régressions non fondées se situent dans la zone III, qui est alors, à tort, considérée comme fiable dans le débat entre significativité substantive et significativité statistique. Il y aura donc des inférences erronées dans la zone III.

Un troisième phénomène établit un lien entre l'inférence à la Fisher et ces corrélations non fondées. Il existe aussi d'autres fonctions de perte que celle du planificateur social bienveillant de la science minimisant les erreurs scientifiques. Il s'agit des fonctions de perte des chercheurs individuels dont la carrière scientifique dépend de la norme des critères de publications.

3. Artefact de publication, méta-analyse et régressions non fondées

a. Artefact de publication et méta-analyse

Dans cette section, nous présentons les conséquences pour les praticiens du critère de Fisher. L'inférence à la Fisher s'impose dans de nombreuses disciplines scientifiques dans les années cinquante et soixante. Lorsque le test de Student donne un résultat qui conduit au rejet de l'hypothèse que la corrélation partielle est nulle, une norme s'impose parmi les éditeurs de revues scientifiques. Ne sont publiés que les résultats rejetant l'hypothèse de nullité du paramètre mesurant l'effet entre deux variables. Ceci crée un artefact statistique relevant du processus de sélection des articles publiés, qui est souvent décrit par le terme « biais de publication ». Il s'agit d'une extension de l'usage du terme « biais » par les statisticiens pour dénommer l'écart entre un paramètre estimé et la « vraie » valeur du paramètre, dans diverses

situations. Les résultats ne rejetant pas l'hypothèse de nullité du paramètre sont parfois appelés résultats « négatifs » (du test), même si le signe du paramètre est positif.

Dès les années 1950, en médecine, les chercheurs décident de faire des « méta-analyses », en compilant des paramètres estimés d'une liaison entre deux variables particulières dans l'ensemble des études disponibles, utilisant différents échantillons, différentes méthodes d'estimation, réalisées par des chercheurs dont les *a priori* ne sont pas nécessairement les mêmes, etc. Ils établissent des moyennes pondérées de ces effets estimés et de leurs dispersions statistiques.

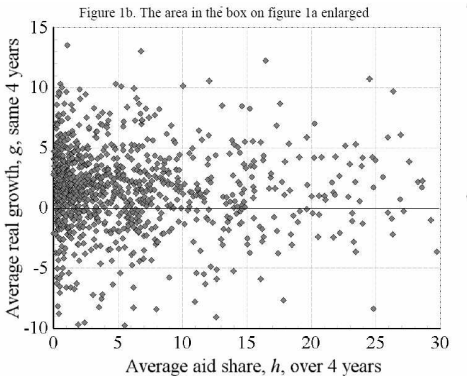
Lors d'une méta-analyse, on peut aussi détecter et corriger l'artefact de publication, parce que l'on a deux dimensions : la taille de l'échantillon et la taille de l'effet pour différentes études. Si les études à petits échantillons ont en moyenne des paramètres plus élevés que les études à grand échantillon, on en déduit que la méthode statistique utilisée par les chercheurs des études à petit échantillon a présenté un artefact à la hausse visant à rendre le paramètre statistiquement significatif (Stanley, 2005). Sur la Figure 2.1, on décèlera un artefact de sélection si l'aire III-L, qui correspond aux effets importants pour des grands échantillons est relativement vide par rapport à l'aire III-S, qui correspond aux effets importants pour des petits échantillons, et l'aire I, qui correspond aux effets de petite taille pour des grands échantillons. Ces trois aires partitionnent la zone critique de rejet de l'hypothèse nulle.

La théorie statistique énonce qu'on ne doit pas obtenir de relation entre les paramètres estimés et le nombre d'observations. En revanche, l'écart-type estimé des paramètres estimés diminue avec le nombre d'observations. En présence d'un artefact de publication avéré, le calcul de l'effet moyen à partir des résultats des différentes études devra sous-pondérer les paramètres des études à petits échantillons par rapport aux paramètres des études à grands échantillons.

À titre d'exemple, nous présentons des résultats d'une méta-analyse concernant la liaison statistique entre l'aide au développement et la croissance économique (Doucouliagos et Paldam, 2009). La Figure 3.1 donne l'ensemble des observations pour de nombreux pays entre ces deux variables. La forme du nuage de point indique qu'une droite de régression linéaire passant à peu près au milieu du nuage serait horizontale. Le coefficient de corrélation simple entre les deux variables

est nul. Les études font se focaliser sur des régressions multiples où les paramètres pourront ne pas être nuls.

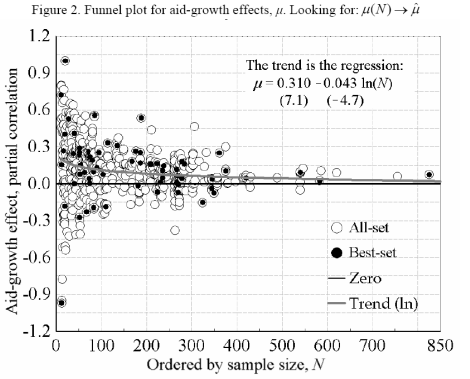
Figure 3.1 : Corrélation simple nulle entre aide au développement de croissance



Source : Doucouliagos et Paldam, 2009, p. 438.

La Figure 3.2 représente cette fois-ci les paramètres estimés dans différentes études en fonction de la taille de l'échantillon de chaque étude. On constate que plus l'échantillon est grand, plus la dispersion des paramètres estimés diminue, en suivant la forme d'un entonnoir. Il s'agit d'un résultat attendu : plus il y a d'observations,

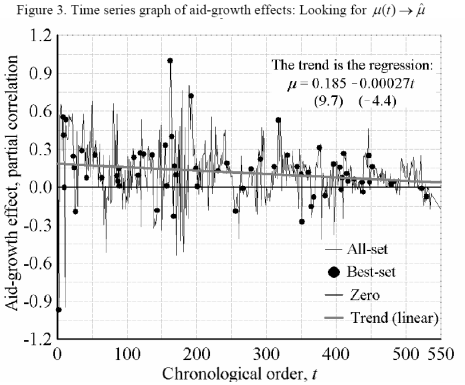
Figure 3.2 : Artefact de publication : le taille de l'effet diminue avec la taille de l'échantillon



Source : Doucouliagos et Paldam, 2009, p. 452.

plus l'écart-type estimé des paramètres estimé est petit. En revanche, on observe qu'il existe une liaison décroissante entre la taille du paramètre estimé et le nombre d'observations. Ce phénomène ne devrait pas apparaître : les paramètres estimés utilisés dans la section 2 n'ont pas de relation avec le nombre d'observations dans les formules données par Yule (1897). Ce résultat inattendu suggère qu'il y a un artefact de publication : les études à petits échantillons ont des paramètres plus élevés.

Figure 3.3 Artefact de publication temporel : la taille de l'effet diminue au cours du temps.



Source : Doucouliagos et Paldam, 2009, p. 452.

La Figure 3.3 montre qu'il existe aussi un artefact de publication temporel : la taille de l'effet publié a tendance à diminuer en fonction de l'ordre chronologique des publications. On a ainsi ce que Ioannidis (2008) associe à un résultat appelé la « malédiction du vainqueur » dans des enchères. Les chercheurs surenchérissent sur la taille d'un effet *nouveau et inattendu* pour obtenir une publication dans une revue scientifique prestigieuse. Ce faisant, il y a des risques importants que les études suivantes qui vont reproduire les résultats dans des revues moins prestigieuses montrent progressivement que l'effet est absent. Ce résultat sera finalement flagrant lors des premières méta-analyses qui apparaîtront plusieurs années après, lorsqu'au moins une trentaine de publications seront disponibles. Ce processus crée des controverses, accroît les citations de l'article initial et la notoriété de ses auteurs. Par un effet de rétroaction positive, ce processus valide *a posteriori*, par les citations, la qualité de la revue, ce qui conduit à renforcer les phénomènes de surenchère sur les

effets nouveaux, inattendus (voire bizarres) et de forte taille dans les publications suivantes.

b. Régressions non fondées, régressions appréciées…

C'est à ce stade que nos régressions non fondées deviennent des régressions appréciées pour les chercheurs recherchant la notoriété. Elles ont quatre avantages :

(1) les paramètres sont très élevés,

(2) ils permettent d'obtenir des effets nouveaux et inattendu dans l'ensemble des effets que la communauté des chercheurs considère *a priori* comme nuls (et, pour cause, puisqu'ils sont vraiment nuls, sauf dans le cas du modèle homéostatique),

(3) la taille du paramètre estimé et son signe sont très sensibles à l'addition ou la suppression de quelques observations atypiques à fort levier (éloignée de la moyenne des observations de la variable explicative). Ces instabilités sur la *taille* et (encore mieux) sur le *signe* de l'effet vont alimenter des controverses, qui accroîtront la notoriété scientifique des auteurs et des revues qui les publient, pendant environ une quinzaine d'années.

(4) *mais surtout, ils peuvent être « statistiquement significatifs », donc publiables*. En effet, l'ensemble des manuels de statistique considère que le problème de corrélation élevée entre les variables explicatives n'est plus un problème lorsque les paramètres estimés sont statistiquement significatifs. Ils proposent fréquemment comme remède d'augmenter la taille de l'échantillon. Le problème est alors réduit au rôle de l'inflation de la variance estimée d'un paramètre estimé, mesurée par le *VIF* « variance inflation factor ».

En revanche, est négligé l'effet sur le paramètre estimé de la corrélation élevée entre les variables explicatives, mesuré par le *PIF* « parameter inflation factor » (Chatelain et Ralf, 2010). *Notre argument central est qu'obtenir la significativité statistique de Fisher pour les paramètres de la régression multiple ne protège pas du caractère infondé des régressions multiples décrites dans la section 2.*

Comment construire une régression fausse, *lorsque la corrélation simple de x_2 sur x_1 est proche de zéro*, comme dans le cas de l'aide au développement et la croissance (cf. Figure 3.1) ? Le Tableau 3.1 propose quatre idées pour trouver une autre variable explicative x_3 très corrélée avec x_2 afin d'obtenir une régression fausse.

Tableau 3.1. Construction de paires de variables explicatives très corrélées entre elles.

x_3 très corrélé avec x_2	Intérêt du modèle :
Indicateurs mesurant deux phénomènes proches ou ayant une cause commune.	Exemple : avis d'experts sur la corruption puis sur le risque d'expropriation dans un pays donné. Recherche de précision dans la différentiation des effets.
Terme retardé : $x_3 = x_2(t-1)$	Modèle dynamique : on distingue les effets de court terme des effets retardés et des effets de long terme.
Puissances : $x_3 = (x_2)^2$ $x_3 = (x_2)^3$	Modèle non linéaire avec des effets marginaux décroissants ou croissants. Approximation polynomiale à l'ordre deux, trois ou quatre de relations non linéaires quelconques.
Terme d'interaction $x_3 = x_2 * x_4$	Complémentarité, modélisation de l'interdépendance échappant à l'hypothèse « toutes autres choses égales par ailleurs ».

Le premier exemple correspond aux cas où il existe plusieurs variables explicatives qui mesurent des phénomènes qui sont associées à des phénomènes très proches, ou lorsqu'un grand nombre de variables explicatives sont disponibles : dans ce cas, on finira bien par obtenir deux variables très corrélées entre elles.

Le second exemple permettra d'obtenir un effet de court terme élevé et par exemple positif, compensé par un effet de la même variable mesuré à la période précédente du signe opposé et élevé. L'effet de long terme faisant la somme des deux effets consécutifs sera quasi nul. On aura dans ce cas fait apparaître un effet de court terme fallacieux.

L'estimation d'un modèle non linéaire est intéressante, car elle permet d'évaluer des effets marginaux croissants et/ou décroissants. Malheureusement, plus la moyenne des observations est éloignée de zéro, plus la corrélation entre une variable et la même variable élevée au carré (ou à une autre puissance) est élevé.

Lorsque la corrélation simple entre x_2 et x_1 est proche de zéro, on pourra obtenir un modèle non-linéaire infondé, dont les paramètres sont statistiquement significatifs.

L'estimation d'un modèle avec terme d'interaction est très riche d'enseignements, car elle prend en compte la complémentarité possible entre deux variables explicatives, et par construction rejette la possibilité d'une intervention « *ceteris paribus* » d'une des deux variables. Par construction, le terme d'interaction est souvent très corrélé avec au moins une des deux variables. Lorsque la corrélation simple entre x_2 et x_1 est proche de zéro, on pourra facilement obtenir un modèle avec terme d'interaction infondé dont les paramètres sont statistiquement significatifs. Il suffira alors d'inventer une narration intéressante autour de cette interaction.

4. La pifométrie au secours de l'économétrie

a. Le PIF et les tests sur les coefficients de corrélation simple

Que peut-on faire pour détecter des résultats issus de régressions non fondées ? Ioannidis (2008) propose de calculer ce qu'il appelle un « ratio de vibration », à savoir le rapport de la taille des effets dans différentes études ou dans différents tableaux statistiques du même article, divisé par le plus petit effet estimé. Un ratio de vibration élevé est un signal d'instabilité de l'effet estimé. Ioannidis (2008) précise également que cette volatilité et cette instabilité de la taille des effets estimés vont être fréquente dans les domaines où l'investigation commence et où la manière d'étudier le phénomène n'est pas définie clairement.

Chatelain et Ralf (2010) proposent le facteur d'inflation du paramètre estimé, ou *PIF* (Parameter Inflation Factor). Il s'agit du rapport entre le paramètre obtenu par la régression multiple divisée par le paramètre obtenu par la régression simple. Si le PIF est supérieur à deux, le paramètre de la régression multiple est égal à plus du double du paramètre de la régression simple. Deux calculs ont déjà été présentés sur l'exemple de la section 2.

À la différence du facteur d'inflation de la variance (*VIF*) qui ne fait intervenir que les coefficients de corrélation simple entre les variables explicatives, le *PIF* fait

aussi intervenir les coefficients de corrélation simple des variables explicatives avec la variable dépendante. En reprenant les notations de Yule (1897), on note r_{12} le coefficient de corrélation simple entre les variables x_1 et x_2. Pour une régression multiple ou la variable dépendante x_1 est expliquée par les deux variables x_2 et x_3 :

(4.1) $$VIF_{32} = \frac{1}{1 - r_{32}^2}$$

et

$$PIF_{12} = \frac{1}{r_{12}} \frac{r_{12} - r_{13} r_{32}}{1 - r_{32}^2} = \left(1 - r_{32} \frac{r_{13}}{r_{12}} \right) \times VIF_{32}$$

Le *VIF* contribue à l'amplification du paramètre mesurée par le *PIF*, en amplifiant l'écart $r_{12} - r_{13}\, r_{32}$. Dans le calcul du paramètre, cet écart correspond, à la contribution de x_2 à l'explication de la variance de x_1 nette de la contribution indirecte de x_2 sur x_1 passant par la médiation de l'autre variable x_3.

Le *PIF* est un outil pour les éditeurs, les rapporteurs et les lecteurs d'articles de revues scientifique utilisant la méthode de régression. Pour le calculer, il faut que les auteurs de l'article présentent, en plus des résultats de leur régression multiple, le nombre d'observations, les moyennes, les écarts-types, et la matrice de corrélation simple de leurs variables. Remarquons qu'une large proportion d'articles omet de présenter la matrice des corrélations, en dépit de la recommandation initiale de Yule (1897).

Pour les auteurs d'articles scientifiques, nous proposons de faire un test de l'hypothèse de nullité du coefficient de corrélation simple entre la variable dépendante et chacune des variables explicatives. On peut aussi faire le test d'une hypothèse composite, telle que le coefficient de corrélation simple ne devrait pas être plus petit que, par exemple, 0,1 (Chatelain et Ralf, 2010). Si ce coefficient s'avère trop petit, on décide de ne pas prendre en compte cette variable explicative de la suite de l'étude. Cette condition préliminaire est incluse dans certaines versions de l'algorithme permettant d'élaborer des graphes causaux en utilisant la méthode de Spirtes *et al.* (2000).

b. Application : Aide au développement, politiques macroéconomiques et croissance économique.

Nous appliquons ces deux outils à un article publié par Burnside et Dollar (2000) dans *American Economic Review*. En dix ans, cet article est devenu le plus cité parmi les articles publiés dans cette revue durant l'année 2000. Pour donner un ordre de grandeur de l'impact de cet article, début octobre 2010, il y avait un peu plus de 2100 citations de cet article référencées dans la base de données d'articles et d'ouvrages académiques utilisée par Google Scholar. Dans cet article, les auteurs montrent que l'aide au développement peut avoir un effet positif sur la croissance seulement s'il y a de bonnes politiques macro-économiques. Qu'entendent-ils par là ? Pas beaucoup d'inflation, un déficit budgétaire faible et une forte ouverture au commerce international. Plus précisément, la variable « politique macroéconomique » est définie par :

$$Politique = 1.28 + 6.85\ surplus\ budgétaire\ de\ l'état - 1.40\ taux\ d'inflation + 2.16\ (exports + imports/PIB)$$

L'implication politique est la suivante : si l'objectif de l'aide au développement est d'augmenter la croissance économique, elle ne devrait être donnée qu'aux pays en développement faisant de « bonnes politiques macroéconomiques ». Leurs résultats sont présentés dans le Tableau 4.1.

Tableau 4.1. Effet de l'aide au développement et des politiques macroéconomiques sur la croissance économique (Burnside et Dollar, 2000), pour N=365 observations.

Aide/PIB	0,034 (0,12)	0,015 (0,012)	0,049 (0,12)
(Aide/PIB) . Politique	-	0,013 (0,049)	0,20* (0,09)
$(Aide/PIB)^2$.Politique	-	-	-0,019* (0,0084)

La variable expliquée est la croissance économique, et le tableau reporte trois des variables explicatives. Le nombre entre parenthèses en dessous du paramètre estimé est l'écart-type estimé du paramètre. Si le ratio de ces deux grandeurs dépasse 1,96, selon le test de Fisher, il existe un effet avec une probabilité de l'erreur de type 1 inférieure à 5% ($p<0.05$). Une habitude est de mettre une étoile lorsqu'un paramètre est « *statistiquement significatif* ». Dans la première colonne, si l'aide apparaît seule, il n'y a pas d'effet statistiquement significatif sur la croissance, comme attendu dans le graphique 3.1. Dans la deuxième colonne, les auteurs utilisent un des outils du Tableau 3.1 : l'ajout d'un terme d'interaction de l'aide avec l'indicateur de politique macroéconomique. Ils n'obtiennent toujours pas de coefficients *« statistiquement significatifs »*. Dans la troisième colonne, les auteurs utilisent un autre outil du Tableau 3.1 : l'ajout d'un terme au carré de l'aide, en interaction avec l'indicateur de politique macroéconomique. Cette fois-ci, les auteurs obtiennent deux paramètres statistiquement significatifs.

Pour calculer les indicateurs que nous venons de proposer, nous avons calculé les coefficients de la régression simple à partir des données téléchargeable sur internet. Nous utilisons l'indice 1 pour la variable dépendante (la croissance économique), l'indice 2 à la variable *(Aide/PIB).Politique*, et l'indice 3 à la variable *(Aide/PIB)*^{2}*.Politique*. Les résultats sont les suivants.

PIF_{12}=0,20/0,095 = 2,13

PIF_{13}=-0,019/0,0046 = -4,15 (avec changement de signe de l'effet).

r_{12}=0,13 : Le test de l'hypothèse $r_{12}=0$ conduit à ne pas la rejeter ($p<0,05$).

r_{13}=0,06 : Le test de l'hypothèse $r_{13}=0$ conduit à ne pas la rejeter ($p<0,05$).

r_{23}=0,92.

On peut calculer le ratio de vibration de Ioannidis pour l'aide/PIB en interaction avec la politique en prenant les paramètres de la deuxième ligne du Tableau 4.1 : *0,20/0,013 = 15,4*. L'ensemble de ces indicateurs confirme qu'il s'agit d'une régression fausse telle que décrite dans la section 1.

Dans l'exemple de la section 1, il était très visible que les paramètres de la paire de variables très corrélées se compensaient (7,08 et -7,01) parce que les deux variables étaient standardisées (elles avaient le même écart-type valant 1). Cette compensation n'est pas décelable dans le Tableau 4.1. En effet, les auteurs d'articles

utilisant la régression présentent généralement les paramètres pour les variables non standardisées :

$$\beta_{12} = \beta_{12}^{S} \frac{\sigma(x_1)}{\sigma(x_2)} = 0{,}20 \text{ et } \beta_{13} = \beta_{13}^{S} \frac{\sigma(x_1)}{\sigma(x_3)} = -0{,}019$$

Du fait du terme au carré pour l'aide, l'écart-type de la variable indicé par 3 est plus grand que celui de la variable indicée par 2. En conséquence, son paramètre non standardisé pourra être nettement plus petit que celui de la variable indicée par 2.

La présentation des coefficients standardisés, programmés dans la plupart des logiciels de statistiques, est donc un autre outil signalant ce problème de corrélation fausse. Dans le cas de la régression simple, le coefficient standardisé est égal le coefficient de corrélation qui est compris entre -1 et 1. En régression multiple, lorsqu'un coefficient standardisé dépasse 1 en valeur absolu, on peut considérer que ce paramètre est sujet à une inflation de sa taille provenant de variables explicatives très corrélées.

L'article de Burnside et Dollar (2000) est un cas exemplaire d'article qui fait face à la « malédiction du vainqueur ». Leur article met en avant un effet très grand et très fragile sur un sujet de politique économique particulièrement sensible : l'aide au développement. Peu après sa publication, il a fait l'objet d'une controverse où Easterly, *et. al.* (2004) ne retrouve pas l'effet de Burnside et Dollar (2000) après l'ajout d'environ quatre-vingts observations. Ensuite, cet article a servi de référence pour un grand nombre d'articles visant à obtenir un effet conditionnel de l'aide au développement sur la croissance, en introduisant d'autres variables explicatives très corrélées avec l'aide, sur le modèle du Tableau 3.1. Enfin, quinze ans après la diffusion du document de travail en 1995, une méta-analyse de Doucouliagos et Paldam (2010) a confirmé l'absence d'effet de l'aide conditionnel à ces variables de politiques économiques sur les études disponibles postérieures à leur article.

Conclusion

Il est simple de résoudre le problème de corrélations non fondées évoqué dans cet article Nous recommandons de ne pas prendre en compte les variables explicatives qui ont des coefficients de corrélation avec la variable dépendante trop petits, et ceci d'autant plus qu'elles sont très corrélées entre elles.

L'obtention de ces régressions non fondées n'est pas intentionnelle, car les arguments que nous avançons dans cet article ne sont pas connus. Elles émergent à l'issue d'un processus évolutionnaire provenant d'une suite d'essais et d'erreurs, afin d'obtenir des paramètres statistiquement significatifs, donc publiables.

Cette suite d'essais et d'erreurs est un autre problème majeur de la validité des résultats des tests d'inférence issus des régressions. Ce problème est aussi appelé celui des *comparaisons multiples*. Le chercheur essaie systématiquement un très grand nombre de variables explicatives, jusqu'à ce qu'émerge un résultat statistiquement significatif. Il suit donc la méthode du professeur Shadoko et du devin plombier pour le lancement de la fusée des Shadoks dans l'espace : « *Ce n'est qu'en essayant continuellement que l'on finit par réussir.* » Les comparaisons multiples successives augmentent la probabilité d'erreur de type I (Denton, 1985). En faisant comme s'il n'y avait pas eu une séquence de comparaisons multiples préliminaires, une inférence utilisant un seuil de 5% pour la probabilité d'erreur de type I est alors erronée. Ce problème conduit lui aussi à des inférences non fondées, qui ne se résument pas aux régressions non fondées traitées dans cet article.

Références

Aldrich, J. (1995), "Correlations Genuine and Spurious in Pearson and Yule", *Statistical Science*, 10(4), pp. 364–76.

Bernard C. [1865] (2008), *Introduction à l'étude de la médecine expérimentale*, (réédition, Collection Champs classiques, Paris : Flammarion).

Burnside C. et D. Dollar (2000), "Aid, Policies and Growth", *American Economic Review*, 90, pp. 847–68.

Bühlmann P., M. Kalisch et M.H. Maathuis (2010), "Variable Selection in High-Dimensional Models: Partially Faithfull Distributions and the PC-Simple Algorithm" *Biometrika*, 97, pp. 261–78.

Chatelain, J.B. et K. Ralf (2010), "Spurious Regressions and Near-Multicollinearity, with an Application to Aid, Policies and Growth".

Cohen, J. (1994), "The earth is round ($p < .05$)", *American Psychologist,* 49(12), pp.997–1003.

Denton, F.T. (1985), "Data Mining as an Industry", *The Review of Economics and Statistics*, 67(1), pp. 124–27.

Doucouliagos, H. et M. Paldam (2009), "The Aid Effectiveness Literature: The Sad Results of 40 Years of Research", *Journal of Economic Surveys*, 23(3), pp. 433–61.

Doucouliagos, H. et M. Paldam (2010), "Conditional Aid Effectiveness: A Meta Study", *Journal of International Development,* 22(4), pp. 391–410.

Easterly, W., R. Levine et D. Roodman (2004), "New Data, New Doubts: A Comment on Burnside and Dollar's 'Aid, Policies, and Growth' (2000)", *American Economic Review* 94(3), pp. 774–80.

Fisher, R. (1925), "Applications of 'Student's' distribution", *Metron*, 5(3), pp. 90–104.

Freedman D. (1997), "From Association to Causation via Regression" in V.R. McKim and S.P. Turner (eds), *Causality in Crisis?*, Notre Dame, IN : University of Notre Dame Press, pp. 113–61.

Galton, F. (1886), "Regression Towards Mediocrity in Hereditary Stature", *The Journal of the Anthropological Institute of Great Britain and Ireland*, 15, pp. 246–63.

Hoover, K.D. (2001), *Causality in Macroeconomics.* Cambridge, UK: Cambrige University Press.

Hume, D. [1739], *A Treatise of Human Nature: Being an Attempt to Introduce the Experimental Method of Reasoning into Moral Subjects*, republié en D.F. Norton et M.J. Norton (eds) (2000), *A Treatise of Human Nature*, Oxford : Oxford University Press.

Ioannidis, J.P.A. (2008), "Why Most Discovered True Associations Are Inflated", *Epidemiology*, 19(5), pp. 640–48.

Keuzenkamp H. (2000), *Probability, Econometrics and Truth. The Methodology of Econometrics*, Cambridge, UK : Cambridge University Press.

Legendre, A.M. (1805), *Nouvelles methodes pour la détermination des orbites des comètes*, Paris : Courcier.

McCloskey, D. et S.T. Ziliak (2008), *The Cult of Statistical Significance: How the Standard Error Cost Us Jobs, Justice and Lives*. Ann Arbor, MI : University of Michigan Press.

Milton J.R. (1987), "Induction before Hume", *The British Journal for the Philosophy of Science*, 38, (3), pp. 49-74.

Moore, H.L. (1905), "The Personality of Antoine Augustin Cournot", *The Quarterly Journal of Economics*, 19, (3), pp. 370-399.

Moore, H.L. (1917), *Forecasting the Yield and the Price of Cotton*, New York : Macmillan.

Pearl, J. (2009), *Causality: Models, Reasoning and Inference* (2nd edition), Cambrige, UK : Cambridge University Press.

Pearson, K. (1897), "On a Form of Spurious Correlation that May Arise when Indices Are Used in the Measurement of Organs", *Proceedings of the Royal Society London Series.* A, 60, pp. 489–98.

Sextus Empiricus [v. 200] (1997), [Πυῤῥώνειοι ὑποτύπωσεις], *Esquisses pyrrhoniennes*, traduction par P. Pellegrin, Paris : Le Seuil.

Simon, H. (1954),. "Spurious Correlation: A Causal Interpretation", *Journal of the American Statistical Association*, 49, pp. 467–92.

Spirtes, P., C.N. Glymour et R. Scheines (2000), *Causation, Prediction, and Search* (2nd edition), Cambridge, UK : Cambrige University Press.

Stanley, T.D. (2005), "Beyond Publication Bias", *Journal of Economic Surveys*, 19, pp. 309–45.

Student (1908), "The Probable Error of a Mean", *Biometrika*, 6, pp. 1–25.

Wiener, N. (1948), *Cybernetics, or Control and Communication in the Animal and the Machine*. Cambridge, MA : MIT Press.

Wright, S. (1920), "The Relative Importance of Heredity and Environment in Determining the Piebald Pattern of Guinea-Pigs", *Proceedings of the National Academy of Sciences*, 6, pp. 320–32.

Yule, G.U. (1897), "On the theory of correlation", *Journal of the Royal Statistical Society*, 60, pp. 812–854.

Réponse de Xavier Ragot (Banque de France)

Je trouve cet article passionnant dans la démarche. Il identifie un problème statistique multiplié par une dynamique sociologique de la profession, qui aboutit à une production erronée de connaissances scientifiques. Il y a donc un artefact global dans la discipline, en particulier dans la mienne, en économie : il y a ce double problème statistique simple parfaitement identifié, ajouté à la sociologie du milieu qui est aussi parfaitement décrite.

Et en plus de cela, Jean-Bernard a la solution, le PIF, pour remettre la science sur la bonne voie. C'est donc un article assez prométhéen dans son ambition, et on a envie ensuite de l'application à grande échelle : de reprendre avec un PIF les méta-analyses déjà faites pour voir l'efficacité du PIF mis en avant comme méthode statistique d'identification. C'est le côté extrêmement enthousiasmant de l'article. Il y a là un champ énorme.

En ce qui concerne la façon dont il est écrit, on peut dire qu'il est très touffu : il contient de l'histoire de l'économétrie, de la sociologie de la science, de la statistique…

Le côté déprimant, c'est que forcément dans la discipline, notamment en économétrie, on sait *a priori* que l'on va identifier de petits effets avec des variables explicatives qui seront très corrélées. On est donc quasiment structurellement dans le cas que tu poses. Quand on prend un pays en voie de développement, tout est corrélé : il n'y a pas de niveau d'éducation, pas d'infrastructure, il y a des maladies, tout est corrélé dans les variables explicatives et on va quand même essayer d'identifier l'aide au développement, parce que l'on a un *a priori* que cela sert à quelque chose, et donc forcément, malheureusement, tu nous dis que même s'il y a un effet, scientifiquement, on ne pourra pas le dire.

Donc finalement, j'en déduis que l'aide au développement aux pays pauvres ne peut pas être justifiée sur une base scientifique étant donné la taille actuelle de l'échantillon que l'on a. Soit on le fait, soit on ne le fait pas, mais tu nous dis que les données ne nous permettent pas de trancher ces questions.

Par rapport à l'économétrie, tu ouvres donc un très grand champ d'indécidabilité des corrélations – je ne parlerai pas des causalités. Il y a une

tentative actuelle, si on prend la précédente médaille Clark, de tout passer, dans les politiques publiques, au filtre de l'évaluation en économétrie : il ne faut faire que ce que l'on a évalué. Et ton article dit que ce n'est pas un programme de recherche prometteur, parce que le champ d'indécidabilité de nos méthodes statistiques est beaucoup trop vaste, et donc même si on corrige des PIF, on ne pourra pas déterminer des causalités. On est donc finalement condamné à la pauvreté statistique des causalités, ce qui est la conclusion pessimiste du statisticien.