Les liaisons fallacieuses : quasi-colinéarité et " suppresseur classique ", aide au développement et croissance

Jean-Bernard Chatelain, Kirsten Ralf

Documents de Travail du
Centre d'Economie de la Sorbonne

CES Working Papers

Les liaisons fallacieuses : quasi-colinéarité et « suppresseur classique », aide au développement et croissance

Jean-Bernard CHATELAIN, Kirsten RALF

2012.11

Maison des Sciences Économiques, 106-112 boulevard de L'Hôpital, 75647 Paris Cedex 13
http://centredeconomiesorbonne.univ-paris1.fr/bandeau-haut/documents-de-travail/
ISSN : 1955-611X

Les liaisons fallacieuses : quasi-colinéarité et « suppresseur classique », aide au développement et croissance

Jean-Bernard Chatelain[1]
Kirsten Ralf[2]

Cet article montre qu'une régression multiple avec deux variables explicatives très corrélées entre elles et dont les corrélations simples avec une variable dépendante sont quasi nulles peuvent correspondre soit à une régression fallacieuse, soit à un modèle homéostatique très sensible à la présence d'observations atypiques à fort levier. La méthode de régression ne permet pas de distinguer entre ces deux modèles. La significativité statistique des paramètres, par ailleurs très élevés, est facile à obtenir dans ce cas, comme le montre une simulation de Monte-Carlo. Un exemple est donné par l'article de Burnside et Dollar [2000], sur les liens entre aide au développement, politiques économiques et croissance.

FALLACIOUS LIAISONS: NEAR MULTICOLLINEARITY AND CLASSICAL SUPPRESSORS, AID, POLICIES AND GROWTH

This paper shows that a multiple regression with two highly correlated explanatory variables, both of them with a near zero correlation with the dependent variable may correspond to a spurious regression or to a homeostatic model, with estimates highly sensible to outliers. The regression method does not allow how to decide which one of the two models is relevant. Statistical significance of the (very high) parameters is easily obtained, as shown doing Monte Carlo simulations. An example is provided by the Burnside and Dollar [2000] article on aid, policies and growth.

Mots-clés : régression fallacieuse, quasi-colinéarité, "suppresseur classique", facteur d'inflation d'un paramètre, aide au développement, croissance économique.

Keywords: spurious regression, near-multicollinearity, classical suppressor, parameter inflation factor (PIF), aid, economic growth

Classification *JEL*: C12, C18, C52, F35, O47

[1] CES (Centre d'Économie de la Sorbonne), Université Paris 1 Panthéon Sorbonne, Paris School of Economics. *Correspondance :* 106-112 boulevard de l'Hôpital, 75647 Paris cedex 13. *Courriel :* jean-bernard.chatelain@univ-paris1.fr
[2] ESCE (École Supérieure du Commerce Extérieur). *Correspondance :* 12 avenue Léonard de Vinci, 92400 Courbevoie. *Courriel :* kirsten.ralf@esce.fr

Les auteurs remercient Michel Armatte, Glenn Shafer, Xavier Ragot, Jean-Philippe Touffut, Christophe Hurlin, Phu Nguyen Van, Giacomo Corneo, Marie Bessec et Christian de Peretti pour leurs commentaires, qui ont permis d'améliorer le contenu de cet article. Ils restent responsables des erreurs et omissions qui pourraient subsister.

INTRODUCTION

Dans une régression multiple, faut-il conserver des variables explicatives dont la corrélation simple avec la variable dépendante est proche de zéro (« classical suppressors » en anglais, qu'on pourrait traduire par « suppresseurs classiques ») ? La pratique usuelle est de les conserver lorsqu'elles sont statistiquement significatives, afin d'éviter un biais de variable omise. Au contraire, cet article recommande de ne pas les intégrer dans les régressions multiples, même lorsqu'elles sont statistiquement significatives.

En premier lieu, l'introduction d'une nouvelle variable explicative dans une régression multiple pourra donner lieu à deux interprétations indécidables à partir des observations statistiques par la méthode de régression (Hoover [2001], p. 45-46). La première interprétation sera que la deuxième variable explicative a un effet de rétroaction négative quasi parfait (et donc une très forte corrélation avec la première variable explicative). Il s'agit d'un modèle « homéostatique », où la corrélation simple quasi nulle est issue d'une stabilisation quasi automatique des chocs provenant de la première variable explicative par une réaction immédiate de la seconde variable explicative. La seconde interprétation est que la première variable n'a aucun effet, même en régression multiple, ce qui peut être montré en « orthogonalisant » les deux variables explicatives. La variance de la variable dépendante est expliquée par la variance du résidu d'une régression intermédiaire où la seconde variable explicative est régressée sur la première variable explicative. Dans ce cas, la variance de la première variable explicative n'explique rien de la variance de la variable dépendante.

S'il y a une forte corrélation entre ces deux variables explicatives, elle va induire un inconvénient supplémentaire (l'*instabilité*) à celui de l'*indécidabilité* de l'interprétation de la régression multiple. Le résidu, issu d'une régression intermédiaire entre les deux variables explicatives très corrélées entre elles, aura une toute petite variance. La régression multiple après orthogonalisation des variables explicatives aura alors un paramètre estimé très élevé et très sensible en termes de signe et de taille du paramètre à la présence d'observations éloignées de la moyenne de ces résidus (aussi appelées « observations à fort levier »).

En second lieu, pour une régression multiple incluant une variable explicative à corrélation simple quasi nulle avec la variable dépendante, l'obtention de la significativité statistique par le test de Student en régression multiple ne sera aisément obtenue que lorsque la corrélation simple entre les deux variables explicatives sera assez élevée (par exemple, de l'ordre de 0,8). Ce résultat provient de la déformation de la zone critique du test de Student. Lorsque deux variables sont très corrélées entre elles, leurs corrélations simples avec une troisième variable sont nécessairement proches. Il suffit que ces deux corrélations simples avec la variable dépendante soient légèrement différentes pour atteindre la significativité statistique, même avec seulement une centaine d'observations. Pour rendre statistiquement significatif un suppresseur classique, il est nécessaire d'introduire une variable explicative très corrélée avec lui.

En troisième lieu, nous expliquons comment l'article extrêmement cité de Burnside et Dollar [2000], montrant que l'aide au développement a un effet sur la croissance économique seulement pour les pays qui ont de bonnes politiques macroéconomiques, est un cas particulier de ces *régressions à effet statistiquement significatif, mais indécidables et instables.* Par exemple, il suffit d'ajouter ou d'enlever trois observations à fort levier (concernant le Botswana) pour faire disparaître ou apparaître leur résultat.

UNE LIAISON PARTIELLE INDÉCIDABLE AVEC DEUX VARIABLES EXPLICATIVES PEU CORRÉLÉES AVEC LA VARIABLE DÉPENDANTE ET TRÈS CORRÉLÉES ENTRE ELLES

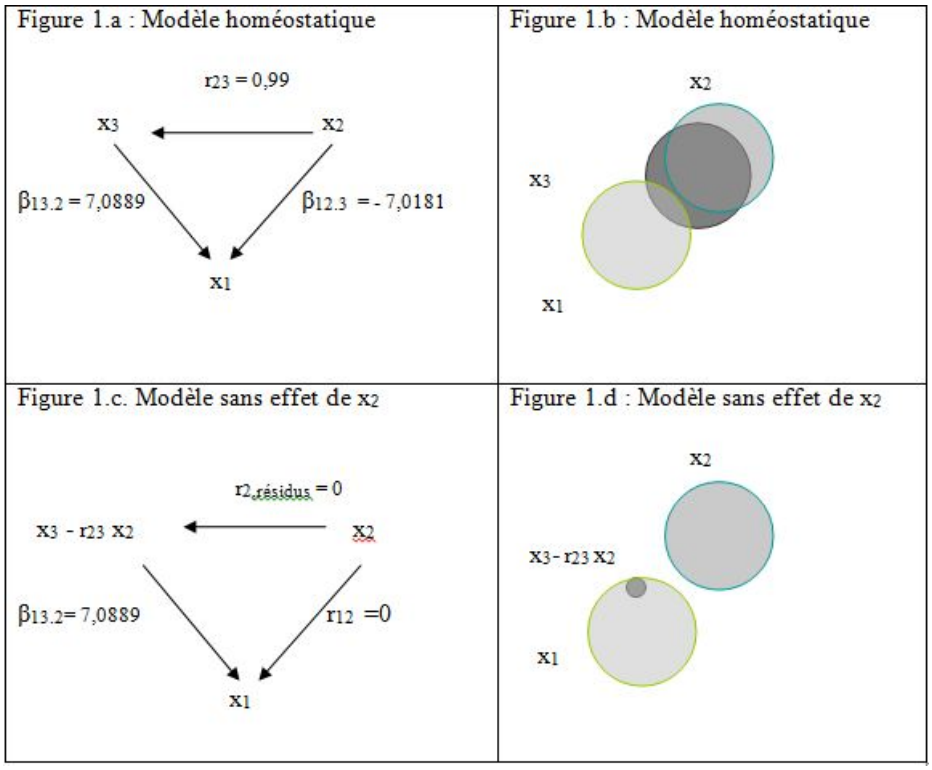
La régression simple estime une relation linéaire entre deux variables observées N fois. Nous considérons des variables « standardisées », de moyenne nulle et d'écart type égal à l'unité. On peut toujours standardiser des variables en soustrayant aux observations leur moyenne et en divisant le tout par leur écart type. 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, avec le coefficient de détermination R^2:

$$x_2 = 0{,}99\, x_3 + \varepsilon_{2.3} \qquad R^2 = 0{,}99^2 = 98\ \%$$
$$x_1 = 0\, x_2 + \varepsilon_{1.2} \qquad R^2 = 0\ \%$$
$$x_1 = 0{,}14107\, x_3 + \varepsilon_{1.3} \qquad R^2 = 2\ \%$$

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 (fig. 1b), 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.

Figures 1a à 1d. *Indécidabilité entre le modèle homéostatique et le modèle sans effet de la variable x_2 après « orthogonalisation » : chemins causaux (1a et 1c) et diagrammes de Venn (1b et 1d)*



Nous notons r_{12} le coefficient de corrélation entre la variable x_1 et la variable x_2 et obtenons les résultats pour la régression multiple suivante (avec des coefficients arrondis à la quatrième décimale) :

(1) $x_1 = -7{,}0181\, x_2 + 7{,}0889\, x_3 + \varepsilon_{1.23}$

$R^2 = -7{,}0181.\, r_{12} + 7{,}0889\, .\, r_{13} = 7{,}0889\, .\, 0{,}14107 = 100\ \%$

$$\beta_{12.3} = \frac{r_{12} - r_{13}r_{23}}{1 - r_{23}^2} = -7.0181 \qquad \beta_{13.2} = \frac{r_{13} - r_{12}r_{23}}{1 - r_{23}^2} = 7.0889$$

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 – 7,0181 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 7,0889 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 [2012] comme une extension du *VIF (Variance Inflation Factor)*. Il est défini comme le 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. Pour une régression multiple ou la variable dépendante x_1 est expliquée par les deux variables x_2 et x_3 :

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

$PIF_{1,2}$ = -7,0181/0 $\qquad\qquad$ $PIF_{1,3}$ = 7,0889/0,14107 =50,25.

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. On peut le calculer pour chacune des variables explicatives de la variable x_1. Pour la variable x_2, le $PIF_{1,2}$ est infini, tandis que pour la variable x_3, le $PIF_{1,3}$ est de 50,25. Les tailles des effets des deux variables x_2 et x_3 sur la variable x_1 ont été considérablement amplifiées.

Une première interprétation du modèle de l'équation 1 est décrite par la figure 1a qui présente l'analyse des chemins des liaisons statistiques « partielles ». Il y a un effet partiel direct de x_2 sur x_1, donné par le coefficient – 7,0181. 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 7,0889. 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 (aux erreurs d'arrondis près) :

$$\beta_{12.3} + r_{13}\, \beta_{13.2} = -7{,}0181 + 0{,}99 \text{ x } 7{,}0889 = 0 = r_{12}$$

La régression multiple décrit un modèle parfaitement « homéostatique ». 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], p. 45-46).

Une seconde interprétation du modèle de l'équation 1 est qu'il s'agit d'une régression fallacieuse où la variable x_2 n'a pas d'effet sur la variable x_1. 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 (fig. 1c) :

$$x_1 = 0\, x_2 + 7{,}0889\, (x_3 - 0{,}99\, x_2) + \varepsilon_{1.23} \qquad\qquad R^2 = 100\ \%$$

La nouvelle régression multiple obtenue correspond à la mise en facteur du paramètre 7,08 de la première régression multiple (équation 1). 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 (fig. 1c) :

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 relativement à celle de la variable dépendante. 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 (fig. 1d)**,** 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 (fig. 1d).

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.

$$7{,}0889 = 1 \times \frac{\sigma(x_1)}{\sigma(x_3 - 0{,}99x_2)} = \frac{1}{\sqrt{1 - r_{23}^2}} = \frac{1}{\sqrt{1 - 0{,}99^2}} = \frac{\mathrm{cov}(x_1, x_3 - 0{,}99x_2)}{\sigma^2(x_3 - 0{,}99x_2)}$$

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 sont 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.

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, plus générale, est beaucoup plus facile à publier.

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.

LES TESTS DE LA STABILITÉ DES INDÉPENDANCES CONDITIONNELLES

Afin d'éliminer le problème d'indécidabilité de la section précédente, Spirtes, *et al.* [2000] et Pearl [2009] font l'hypothèse, *a priori*, que les violations de la *« condition de stabilité des indépendances conditionnelles »* (un paramètre serait nul en régression simple et différent de zéro en régression multiple, et vice-versa) sont très rares. Leur argument est qu'une égalité stricte entre paramètres serait de mesure nulle (telle que : 0,99 x 7,08 – 7,01 = 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). En revanche, Freedman [1997] avance qu'on peut tout à fait accepter l'hypothèse *a priori* alternative : il n'y a pas de raison de rejeter l'hypothèse que des paramètres soient liés par des contraintes d'égalité.

Dans la pratique des économètres, la stabilité des indépendances conditionnelles peut être testée à l'aide de deux tests de Student sur des échantillons d'observations. Chatelain et Ralf [2012] montrent que l'obtention de la significativité statistique par le test de Student en régression multiple ne sera aisément obtenue que lorsque la corrélation simple entre les deux variables explicatives sera assez élevée (par exemple, de l'ordre de 0,8), pour une régression multiple incluant une variable explicative à corrélation simple quasi nulle avec la variable dépendante. Ce résultat provient de la déformation de la zone critique du test de Student lorsque deux variables sont très corrélées entre elles. Dans ce cas, leurs corrélations possibles avec une troisième variable sont nécessairement proches. Il suffit alors que ces deux dernières corrélations soient légèrement différentes pour des petits échantillons pour atteindre la significativité statistique.

Nous effectuons des simulations de Monte-Carlo de tirages d'échantillons de loi multi-normales de moyennes nulles et d'écarts types égaux à l'unité. La variable x_2 a une corrélation théorique nulle r_{12} avec la variable x_1. Les simulations par ordinateur de tirages d'échantillons aléatoires, dont les valeurs numériques sont calculées à la douzième décimale près, sont telles que jamais on obtient une corrélation exactement égale à zéro. En revanche, pour chaque échantillon, on peut faire des tests d'inférence sur l'hypothèse $r_{12} = 0$ et sur l'hypothèse $\beta_{12.3} = 0$. On peut ensuite faire calculer les pourcentages de stabilité ou d'instabilité des indépendances conditionnelles. Elles sont reportées dans le tableau 1.

On constate que, lorsque la corrélation entre les variables explicatives est forte mais pas très élevée ($r_{23} = 0{,}50$), les tests concluent à la stabilité de l'indépendance conditionnelle associée aux hypothèses d'absence d'effets de la variable x_2 pour plus de 90 % des échantillons. De surcroît, ce pourcentage varie peu avec l'accroissement du nombre d'observations de chaque échantillon (il passe de 92,3 % pour 102 observations à 90,6 % pour 1 002 observations).

En revanche, lorsque la corrélation entre les variables explicatives est très élevée ($r_{23} = 0{,}99$), les tests concluent à la stabilité de l'indépendance conditionnelle associée aux hypothèses d'absence d'effets de la variable x_2 pour seulement 45,3,% des échantillons de 102 observations et pour 0 % des échantillons de 1 002 observations. Ceci est associé à 52,5 % de cas d'inférence indécidable pour les échantillons de 102 observations puis de 95,5 % de cas d'inférence indécidable pour les échantillons de 1 002 observations.

Une recommandation qui consisterait à accroître le nombre d'observations favoriserait donc des inférences d'*effets statistiquement significatifs qui seraient en réalité indécidables et instables*, lorsque la corrélation entre les variables explicatives est très élevée et lorsque les corrélations entre ces variables explicatives et la variable dépendante sont très faibles.

Tableau 1. *Inférences sur la stabilité des indépendances conditionnelles pour 1 000 échantillons aléatoires de lois multi-normales à corrélation très faible avec la variable explicatives $r_{12} = 0$, $r_{13} = -0{,}03$, en faisant varier la corrélation entre les variables explicatives r_{23} et le nombre N d'observations des échantillons.*

$r_{23} = 0{,}50$, N = 102 observations	Ne rejette pas $r_{12} = 0$	Rejette $r_{12} = 0$
Ne rejette pas $\beta_{12.3} = 0$	92,3 %	3,3 %
Rejette $\beta_{12.3} = 0$	2,1 % (« indécidable »)	2,3 %

$r_{23} = 0{,}50$, N = 1 002 observations	Ne rejette pas $r_{12} = 0$	Rejette $r_{12} = 0$
Ne rejette pas $\beta_{12.3} = 0$	90,6 %	2,4 %
Rejette $\beta_{12.3} = 0$	4,9 % (« indécidable »)	2,1 %

$r_{23} = 0{,}99$, N =102 observations	Ne rejette pas $r_{12} = 0$	Rejette $r_{12} = 0$
Ne rejette pas $\beta_{12.3} = 0$	42,3 %	2,8%
Rejette $\beta_{12.3} = 0$	52,3 % (« indécidable »)	2,8%

$r_{23} = 0{,}99$, N =1 002 observations	Ne rejette pas $r_{12} = 0$	Rejette $r_{12} = 0$
Ne rejette pas $\beta_{12.3} = 0$	0 %	0 %
Rejette $\beta_{12.3} = 0$	95,5 % (« indécidable »)	4,5 %

APPLICATION : AIDE AU DÉVELOPPEMENT, POLITIQUES MACROÉCONOMIQUES ET CROISSANCE ÉCONOMIQUE

L'article publié par Burnside et Dollar [2000] dans *American Economic Review* est devenu le plus cité parmi les articles publiés dans cette revue durant l'année 2000. Début février 2012, il y avait un peu plus de 2 780 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 macroéconomiques. 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 (sans les effets des variables de contrôle) dans le tableau 2.

Tableau 2. *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 % ($p < 0{,}05$). Un usage 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. Dans la deuxième colonne, les auteurs ajoutent 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 ajoutent 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.

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*. Nous obtenons les résultats suivants qui suggèrent la présence d'une inférence indécidable :

$PIF_{12} = 0{,}20/0{,}095 = 2{,}13$
$PIF_{13} = -0{,}019/0{,}0046 = -4{,}15$ (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 [2008] pour l'aide/PIB en interaction avec la politique en prenant les paramètres de la deuxième ligne du tableau 2 : 0,20/0,013 = 15,4. Il s'agit du rapport de la taille des effets dans différentes études ou dans différents tableaux statistiques du même article, divisé par le plus petit effet estimé. Comme le *PIF*, un ratio de vibration élevé est un signal d'instabilité de l'effet estimé.

L'ensemble de ces indicateurs confirme qu'il s'agit d'une inférence indécidable. 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 2. 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.

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 retrouvent 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. 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.

Cet exemple et les travaux de Ioannidis [2008] en épidémiologie suggèrent que ces régressions multiples *« à effets statistiquement significatifs mais indécidables et instables »* pourraient présenter des opportunités fréquentes favorisant la carrière des chercheurs compte tenu des biais de publications des revues scientifiques prestigieuses. Elles permettent d'obtenir des effets nouveaux et inattendus (et pour cause, puisqu'il peut ne pas y avoir d'effet dans le vrai modèle) particulièrement prisées par ces revues, statistiquement significatifs et très instables en termes de signe et de taille de l'effet. Ce dernier point favorise les controverses, les citations afférentes et la notoriété des chercheurs et des revues publiant ce genre de régression. Elles sont également faciles à construire et adossées à des modèles théoriques intéressants : il suffit de trouver une deuxième variable explicative très corrélée au suppresseur classique. Ce peut être une variable qui a un facteur causal non observé identique à la première variable explicative, un terme retardé pour un modèle dynamique, un terme au carré pour un modèle non linéaire, un terme d'interaction pour un modèle rejetant l'hypothèse *ceteris paribus* entre deux variables explicatives.

Ces régressions indécidables pourraient donc être surreprésentées dans les revues prestigieuses. Elles orienteraient l'accumulation des recherches dans une mauvaise direction pendant une quinzaine d'années, jusqu'à ce que des méta-analyses conduisent à abandonner l'hypothèse testée. On observerait alors un phénomène proche de la « malédiction du vainqueur » dans des enchères ou surenchères scientifiques. En conduisant à une mauvaise allocation des ressources des chercheurs, ces régressions *à effet statistiquement significatif mais indécidable et instable* pourraient être à l'origine d'un coût social très élevé pour l'activité scientifique lorsqu'elle utilise l'inférence en régression multiple comme élément de preuve. Ces liaisons fallacieuses seraient alors autant de liaisons dangereuses pour la science.

CONCLUSION

Pour éviter le problème de ces régressions *à effet élevé (en valeur absolue) et statistiquement significatif mais indécidable et instable* évoqué dans cet article, il suffit de ne pas prendre en compte les variables explicatives qui ont des coefficients de corrélation simple avec la variable dépendante trop proches de zéro, et ceci d'autant plus que ces variables sont très corrélées entre elles.

RÉFÉRENCES BIBLIOGRAPHIQUES

BURNSIDE C. et DOLLAR D. [2000], « Aid, Policies and Growth », *American Economic Review*, 90, p. 847-868.

CHATELAIN J.-B. et RALF K. [2012], « Spurious Regressions and Near-Multicollinearity, with an Application to Aid, Policies and Growth », *Mimeo*.

DOUCOULIAGOS, H. et PALDAM M. [2010], « Conditional Aid Effectiveness: A Meta Study », *Journal of International Development,* 22 (4), p. 391-410.

EASTERLY W., LEVINE R. et ROODMAN D. [2004], « New Data, New Doubts: A Comment on Burnside and Dollar's 'Aid, Policies, and Growth' (2000) », *American Economic Review*, 94 (3), p. 774-780.

FREEDMAN D. [1997], « From Association to Causation *via* Regression » dans V.R. McKIM et S.P. TURNER (eds), *Causality in Crisis*, p. 113-161, University of Notre Dame Press, Notre Dame.

HOOVER K.D. [2001], *Causality in Macroeconomics*. Cambridge (UK), Cambridge University Press.

IOANNIDIS J.P.A. [2008], « Why Most Discovered True Associations Are Inflated », *Epidemiology*, 19 (5), p. 640-648.

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

SPIRTES P., GLYMOUR C.N. et SCHEINES R. [2000], *Causation, Prediction, and Search*, 2e éd., Cambridge (UK), Cambridge University Press.