Document de travail de la série

Etudes et Documents

E 2008. 25

L'utilisation de variables explicatives estimées dans les régressions économétriques

Michaël Goujon

CERDI-CNRS, Université d'Auvergne

m.goujon@u-clermont1.fr

Cette version : février 2008.

12 p.

Résumé

Cette note a pour objet de rappeler certains résultats de la littérature sur les problèmes posés par l'introduction dans une régression économétrique de variables explicatives qui ont été précédemment estimées à partir d'une équation auxiliaire en première étape. Les indicateurs de politique économique « révélée » par les résultats, telles qu'elles ont été initiées au CERDI il y a déjà un certain nombre d'années font en effet partie de cette classe de variables et ont éventuellement pour finalité d'être utilisés comme variables explicatives dans des régressions économétriques. Cette note présente les principaux résultats de l'article de Pagan (1984) qui est l'article de référence en la matière et des illustrations à partir d'articles utilisant ces résultats.

Introduction

Cette note a pour objet de rappeler certains résultats de la littérature sur les problèmes posés par l'introduction dans une régression économétrique de variables explicatives qui ont été précédemment estimées à partir d'une équation auxiliaire en première étape. Les indicateurs de politique économique « révélée » par les résultats, telles qu'elles ont été initiées au CERDI il y a déjà un certain nombre d'années (Guillaumont et Guillaumont Jeanneney, 1988, Guillaumont et Lefort, 1993, Combes et al., 2000, Brun et al., 2005 et 2007, Boussichas et Goujon, 2008) font en effet partie de cette classe de variables. Ces indicateurs ont éventuellement pour finalité d'être utilisés comme variables explicatives dans des régressions économétriques portant sur les déterminants de résultats économiques. Il convient donc d'en vérifier les modalités d'utilisation dans de tels travaux économétriques.

Ces variables explicatives estimées sont généralement qualifiées de « constructed variables » ou plus souvent de « generated regressors » (GR). L'équation principale où sont introduits les GR, est notée Equation (1) tandis que l'équation auxiliaire d'où sont issus les GR est notée Equation (2) dans la suite.

Les auteurs utilisant les GR considèrent généralement que les estimations des coefficients de l'équation (1) convergent vers les vraies valeurs des coefficients si les estimations convergent également dans l'équation (2). Les estimations des paramètres seraient donc « consistantes ».
En revanche, l'« inférence statistique » basée sur des tests statistiques reposant sur la précision (variance ou écart-type) des paramètres estimés poserait problème. En effet, les auteurs considèrent généralement que les écarts-types et les statistiques de tests obtenus d'une estimation simple en moindres carrés ordinaires (MCO) de l'équation (1) ne sont pas valides car ils ignoreraient la variation d'échantillonnage des paramètres estimés de l'équation (2). En d'autres termes, puisque les paramètres de l'équation (2) sont estimés à partir de données[1], l'incertitude issue de cette estimation devrait être prise en compte dans l'estimation de (1), ce qui n'est pas fait quand on applique les MCO. Il conviendrait alors d'adopter, par exemple, une méthode de correction des écarts-types (le bootstrap par exemple).
Or, cet argument, bien qu'il paraisse logique, n'a pas une portée générale. L'article de Pagan (1984) est l'article de référence sur le sujet et nous en exposons les principaux résultats dans

[1] Même si ce ne sont pas les mêmes données utilisées pour l'estimation des paramètres de l'équation (1).

la première section. Des illustrations sont ensuite présentées dans la deuxième section à partir d'articles faisant référence aux résultats présentés par Pagan (1984).

I – Le traitement des GR dans les régressions économétriques

Pagan (1984) a été le premier à proposer un exposé détaillé des problèmes économétriques posés par l'utilisation de GR dans une régression économétrique et reste une référence en la matière. Pagan distingue deux principaux types de GR :

- ceux basés sur la valeur prédite de la variable expliquée de l'équation (2) (« Predictor-generated regressors » ou PGR).
- ceux basés sur les résidus de l'équation (2) (« Residual-generated regressors » ou RGR)

Nous nous limitons ici à l'exposé des résultats quand est utilisé un RGR dans l'équation (1) puisqu'il correspond au cas des indicateurs de politique économique révélée (voir Annexe pour le cas où seul un PGR est introduit dans l'équation (1)). Les démonstrations de Pagan (1984) concernant l'utilisation des RGR sont basées sur le modèle suivant qui inclut à la fois un RGR et un PGR (« modèle 4 », section 3, p. 232-235):

$$(1)\ y = \delta z^* + \gamma\,(z - z^*) + e$$

$$(2)\ z = z^* + \eta = W\alpha + \eta$$

avec e et η des erreurs normales et indépendantes, de moyennes nulles et de variances respectives σ^2_e et σ^2_η. Dans le cas général, $E(e, \eta)=0$ est une condition nécessaire pour que l'estimation MCO soit consistante. La variable z* est inobservée mais est une fonction des variables observées prédéterminées appartenant au vecteur W.

Pagan (1984) illustre ce modèle 4 en considérant que le terme PGR z* peut, par exemple, représenter la part « anticipée » et le terme RGR (z – z*) la part « non anticipée » (ou de « surprise ») de la variable z (avec l'apparition de la Nouvelle Economie Classique dans les années 1970, différents modèles théoriques ont considéré par exemple que la part anticipée et la part non anticipée de l'inflation ou de la monnaie ont des impacts différents sur l'activité ou le taux de chômage).

Il s'agit bien du cas correspondant aux indicateurs de politique économique révélée, avec le terme RGR ($z - z^*$) qui correspond aux résidus d'une équation explicative d'une variable de résultat z sur des facteurs structurels (W). Le terme PGR z^* correspond alors à la part de la variable de résultat expliquée par les facteurs structurels (qui n'est pas systématiquement introduit dans l'équation (1) dans les travaux dans ce domaine ou, tout du moins, ne constitue pas l'objet principal de l'étude). Le terme RGR ($z - z^*$) capte l'impact de la politique sur le résultat z et constitue l'indicateur de politique révélée. L'équation (1) inclut donc le test de l'impact de la politique sur y, une variable économique, portant sur la nullité du coefficient γ.

Le modèle 4 de Pagan est plus général que le modèle où seul un PGR (z*) apparaît dans (1), le « modèle 1 », qui fait l'objet d'un rappel en Annexe de cette note. Les résultats et préconisations présentés ci-dessous concernant le modèle 4 diffèrent de ceux du modèle 1 présenté en Annexe.

A partir du système composé des équations (1) et (2), on peut obtenir une estimation des coefficients $\hat{\delta}$ et $\hat{\gamma}$:

- soit en utilisant l'estimateur du maximum de vraisemblance (estimation jointe des deux équations),
- soit en utilisant une procédure en deux étapes qui consiste à estimer – typiquement par les MCO – l'équation (2) et notamment le paramètre $\hat{\alpha}$, puis construire $\hat{z} = W\hat{\alpha}$, puis ($\hat{z} - z$), et les introduire dans l'estimation de l'équation (1) pour obtenir $\hat{\delta}$ et $\hat{\gamma}$.

Le Théorème 7 (p.232-233) de Pagan (1984) porte sur la consistance des estimations à la fois des paramètres et des écarts-types :

- L'estimation des paramètres $\hat{\delta}$ et $\hat{\gamma}$ par les MCO appliqués aux deux étapes est efficiente asymptotiquement (les estimations de $\hat{\delta}$ et $\hat{\gamma}$ sont alors égales à celles qui seraient issues d'une estimation par le maximum de vraisemblance du système (1) et (2)).
- L'estimation par les MCO de la variance de $\hat{\delta}$ est incorrecte (non consistante), mais celle de $\hat{\gamma}$ est en revanche correcte (du fait de la convergence de l'estimateur de la variance des résidus vers la vrai variance σ^2_e). Ainsi, pour la procédure à deux étapes appliquée à (1)-(2), Pagan (1984) propose que les valeurs correctes des écarts-types des paramètres devraient être

retirées d'une estimation MCO pour $\hat{\gamma}$ et d'une estimation par l'estimateur des variables instrumentales (EVI) ou des doubles-moindres carrés (DMC) pour $\hat{\delta}$. [2]
Ces propositions générales admettent cependant des cas particuliers.

Si l'hypothèse δ=0 (nullité du paramètre attaché au PGR) est la seule hypothèse que l'on cherche à tester, Pagan (1984) montre que les écarts-types estimés par les MCO ne sont pas corrects mais sont sous-estimés par rapport aux écarts-types corrects. Dans ce cas les t-statistiques sont sur-estimés de telle sorte que l'acceptation de l'hypothèse δ=0 avec les écarts-types MCO ne serait de toute façon pas renversée si les écarts-types corrects avaient été utilisés.

Dans le cas où seul le terme RGR ($\hat{z}$ – z) apparait dans (1), c'est-à-dire quand δ=0, on a le modèle plus simple :

$$(1)\ y = \gamma\,(z - z^*) + e$$

$$(2)\ z = z^* + \eta = W\alpha + \eta$$

Alors, il peut être retiré du Théorème 7 que l'estimation par les MCO de γ et de sa variance est consistante ou correcte[3] (Voir également Pagan, 1986, page 525, proposition 3.3).
Ce cas est important car il correspond à l'utilisation que l'on fait généralement des indicateurs de politique économique révélée, sans introduction du terme PGR qui n'est pas l'objet principal de l'étude (c'est également le cas quand on considère que seules les variables « non anticipées » ont un impact sur y, comme dans Sargent, 1976). De plus, cette conclusion est indépendante de l'introduction de variables explicatives supplémentaires dans (1) ; si ces variables supplémentaires apparaissent dans le vecteur W, alors l'estimateur à deux étapes est parfaitement efficient.

[2] Bohn et Deacon, 2000, utilisent des résidus d'une équation d'investissement sur des variables structurelles comme indicateurs de sécurité de la propriété dans une équation explicative de l'exploration pétrolière. Ils considèrent que les coefficients MCO et les écarts-types pour les régresseurs estimés sont consistants sous l'hypothèse nulle que ces coefficients sont nuls. Les t-statistiques peuvent donc être utilisés pour tester l'absence d'effet. Les écart-types sont en revanche inconsistants sous d'autres hypothèses nulles. Pour obtenir des écarts-types consistants, on peut ré-estimer l'équation principale en utilisant une procédure similaire à l'estimateur des moindres carrés à deux étapes (two-stage least squares). Le résidu est remplacé par la variable expliqué (investissement) et les variables structurelles explicatives de l'équation secondaire. Puisque la variable expliqué de l'équation secondaire est endogène, le modèle est estimé par la méthode des variables instrumentales en utilisant le résidu comme instrument (la variable expliqué prenant le rôle du résidu). Les auteurs observent que les écarts-types sont alors diminués, ce qui renforce la conclusion que les résidus sont explicatifs dans l'équation principale.
[3] Alors qu'une estimation avec la méthode des variables instrumentales – telle qu'elle est préconisée dans le cas où seul un PGR intervient dans (1) – voir annexe – ne seraient pas consistants.

Dans le cas où, à côté des variables courantes, les variables retardées z^*_{t-1} et $(z_{t-1} - z^*_{t-1})$ apparaissent dans (1). Il s'agit d'un modèle différent, utilisé par exemple par Barro (1977), et qui fait l'objet d'une discussion particulière par Pagan (1984), p. 233-234, Théorème 8 relatif au « modèle 5 ».

La méthode à deux étapes pour estimer ce modèle consiste alors à :

- en première étape, régresser z contre W pour obtenir $\hat{z} = W\hat{\alpha}_{-1}$ et $\hat{z}_{-1} = W_{-1}\hat{\alpha}$,
- et en deuxième étape régresser y contre $\hat{z}$, $\hat{z}_{-1}$, $(z - \hat{z})$, et $(z_{-1} - \hat{z}_{-1})$.

Pagan (1984) montre dans le Théorème 8 que la matrice des variances-covariances des erreurs n'est pas sphérique et par conséquent, l'estimation par les MCO des écarts-types des paramètres de l'équation (1) ne peut être correcte. Cependant, là encore, les écarts-types de l'estimation MCO ne sont pas plus grands que les écarts-types corrects. Les t-statistiques sont alors sur-estimés de telle sorte que l'acceptation de l'hypothèse $\delta=0$ pour les termes PGR courants et retardés avec les écarts-types MCO ne serait de toute façon pas renversée si les écarts-types corrects avaient été utilisés [on retrouve la conclusion précédente relative à l'utilisation du seul RGR courants].

En résumé, les résultats présentés par Pagan (1984) sont les suivants. Les MCO appliqués en deux étapes sur le système d'équations où seul un RGR apparaît dans l'équation (1) donnent une estimation consistante des paramètres et des écarts-types d'estimations. Cette conclusion tient quelques soient les autres variables explicatives introduites dans les deux équations, sauf dans les cas où sont introduits un PGR ou un RGR / PGR retardé qui nécessitent un traitement spécifique (certaines hypothèses cependant, notamment celle de nullité des paramètres des PGR, restant testables).

II - Quelques applications

La référence au théorème 7 de Pagan (1984) relatif au traitement des RGR est faite dans des articles traitant de sujets divers et répond à des problèmes techniques qui paraissent différents initialement. Sans que la liste soit exhaustive, nous en présentons quelques exemples parmi les plus significatifs[4].

[4] Kearney (2001) évoque également le théorème de Pagan (1984). L'auteur assimile un paramètre d'une équation (2) variant dans le temps (estimé par le filtre de Kalman) à un RGR et considère que son utilisation dans une équation (1) estimée par les MCO produit des valeurs corrects des paramètres et des estimations consistantes des écarts-types (voir aussi Kearney, 1996).

Gomanee et al (2005) étudient l'impact de l'aide sur la croissance en Afrique Sub-Saharienne en décomposant les canaux de transmission de l'aide en utilisant la technique des GR. Cette technique permet à la fois de prendre en compte l'impact de l'aide direct et à travers des variables intermédiaires (comme l'investissement) et d'éviter un problème de double compte (si on introduisait à la fois l'aide et l'investissement, puisqu'une partie de l'investissement est financée par l'aide). Les auteurs introduisent dans (1), une équation de croissance, un RGR issu de (2), la régression de l'investissement sur l'aide. Les auteurs considèrent que la procédure en deux étapes (utilisant le résultat de (2) dans (1)) avec les MCO donne des estimations des paramètres efficientes asymptotiquement et des valeurs correctes de la variance et des écart-types de ces coefficients. Cette conclusion tient quelques soient les variables explicatives supplémentaires dans les équations (1) ou (2), comme dans leur cas, la variable d'aide qui apparaît dans les deux équations.

McCallum (1987) dans une équation (1), explicative des niveaux de chômage au Canada et aux Etats-Unis, introduit la « part exogène » du taux d'investissement (résidus d'une régression (2) du taux d'investissement sur d'autres variables explicatives) et la « part exogène » du taux d'emploi dans le secteur manufacturier (résidus d'une régression (2') du taux d'emploi dans le secteur manufacturier sur d'autres variables explicatives du modèle). L'auteur considère que les coefficients estimés des autres variables explicatives du modèle ne sont pas affectés par l'emploi de la méthode (voir aussi la démonstration dans Gomanee et al., 2005) ; De plus, l'auteur en référence à Pagan (1984) considère que les estimations MCO de l'équation (1) sont correctes, puisque l'équation ne contient que des RGR et non des PGR ou des RGR / PGR retardés.

Marchetti et Nucci (2005, 2007) teste dans une équation (1) l'impact des chocs technologiques sur le taux d'emploi ; Les chocs technologiques sont les résidus d'un processus autorégressif des résidus d'une équation (2) explicative de la production (au niveau entreprises). Les auteurs, en référence à Pagan (1984), considèrent que l'utilisation de résidus estimés courants n'affecte pas la consistance et l'efficience des estimateurs et la validité de l'inférence statistique standard (ce qui ne serait pas le cas si avaient été introduits des RGR retardés).

Bénassy-Quéré et al. (2007) étudient l'impact des institutions sur les flux d'investissement direct étranger (équation (1)). La méthodologie économétrique prend en compte le fait que la qualité des institutions est affectée par le niveau du PIB par tête qui est également un déterminant des flux d'IDE (il y a donc un problème de collinéarité entre certaines des variables explicatives dans (1)) ; dans ce cas précis, omettre le PIB par tête dans (1) pourrait conduire à une erreur due à une corrélation éventuelle des institutions et du PIB par tête : un coefficient positif des institutions sur l'IDE pourrait en fait capter l'impact non modélisé du PIB par tête sur l'IDE (si le PIB par tête détermine en partie le niveau des institutions). Il s'agit donc, reprenant les termes des auteurs, d' « orthogonaliser » les variables institutionnelles, en les régressant sur le niveau du PIB par tête (équation (2)) ; le résidus de l'équation, un RGR, mesure alors la qualité des institutions indépendante du niveau du PIB par tête. Les auteurs considèrent que l'utilisation d'un RGR ne biaise pas l'estimation de (1) par les MCO ni les écarts-types. Cependant, les auteurs doivent instrumenter le RGR-institutions afin de prendre en compte le fait que les institutions peuvent être endogènes vis-à-vis des flux d'IDE (la variable expliquée de (1)). En conséquence, dans l'équation (1) sont introduites les <u>valeurs prédites</u> des institutions par l'équation d'instrumentation, ce qui nécessite l'utilisation de la technique bootstrap pour les résidus de l'équation d'instrumentation.

De Santis et al. (2004), étudient dans l'équation (1) les déterminants des investissements européens aux Etats-Unis en utilisant le modèle d'investissement Q de Tobin. Leur variable d'intérêt sont les développements sur les marchés boursiers européens, ajustés pour l'évolution économique commune aux deux zones. Cette variable se présente comme le résidu d'une régression (2) des indices de marchés européens sur l'indice de marché américain. Les auteurs évoquent le problème sur la validité de l'inférence statistique posé par les GR, du fait que l'estimation MCO ne traite pas l'incertitude introduite par le GR. Ils rappellent cependant que si ce problème est réel quand est introduit un PGR issu de l'équation (2), Pagan (1984) montre que ce n'est pas le cas pour les RGR. Les MCO donnent alors des estimations consistantes des coefficients et des écarts-types en présence de RGR courants. Comme c'est le cas de leur variable d'intérêt, les auteurs considèrent qu'il n'y a alors pas besoin d'ajuster les écarts-types pour prendre en compte la présence de GR.

Références

Bénassy-Quéré A., Coupet M. et Mayer T. (2007), Institutional Determinants of Foreign Direct Investment, The World Economy, 30, 5, p. 764-782.

Bohn H et Deacon R.T. (2000), Ownership risk, investment and the use of natural resources, The American Economic Review, 90, 3, p. 526-549.

Boussichas M. et Goujon M. (2008), Un indicateur de politique d'ouverture « révélée » des pays OCDE à la migration du Sud, Etudes et Documents du CERDI 2008/06.

Brun J.-F., Chambas G., et Combes J.-L. (2005), Quel niveau de ressources publiques en Afrique sub-saharienne? in Afrique au sud du Sahara : mobiliser des ressources sur le développement, ed G.Chambas, Economica, Paris.

Brun J.-F., Chambas G., et Guérineau S. (2007), Aide et mobilisation fiscale dans les pays en développement, Rapport Thématique Jumbo 21, AfD Département de la Recherche, Paris.

Combes J-L. Guillaumont P. Guillaumont Jeanneney S.et Motel Combes P. (2000), Ouverture sur l'extérieur et instabilité des taux de croissance, *Revue Française d'Economie*, 15, 1, Hiver.

De Santis R. A., Anderton R. et Hijzen A. (2004), On the Determinants of Euro Area FDI to the United States: The Knowledge-Capital-Tobin's Q Framework , ECB Working Paper No. 329.

Dufour J.-M. et Jasiak J. (2000), Finite sample inference methods for simultaneous equations and models with unobserved and generated regressors, CIRANO Working Paper 2000/13 Université du Québec.

Gomanee K, Girma S., et Morrissey O. (2005), Aid and growth in Sub-Saharan Africa: accounting for transmission mechanisms, Journal of International Development, 17, 8, p.1055-1075.

Gomanee K, Morrissey O., et Mosley P. (2005), Aid and Growth , Government expenditure and aggregate welfare, World Development, 33, 3, p.355-370.

Guillaumont P. et Guillaumont Jeanneney S., 1988, (dir) *Stratégies de développement comparées, zone franc et hors zone franc*, Paris, Economica.

Guillaumont P. et Lefort C., 1993, Facteurs structurels et facteurs politiques de l'urbanisation : hypothèses pour les années quatre-vingt, in *Croissance démographique et urbanisation. Politiques de peuplement et aménagement du territoire,* Cahiers de l'AIDELF n°5, pp.275-281.

Kearney A.A. (1996), The effect of changing monetary policy regimes on stock prices, Journal of Macroeconomics, 18, 3, p. 429-447.

Kearney A.A. (2001), A note on modelling the impact of economic announcements on interest rates, Economic Letters 71, p. 83-89.

Marchetti D.J. et Nucci F. (2005), Price Stickiness and the contractionary effect of technology shocks, European Economic Review 49, p.1137-1163.

Marchetti D.J. et Nucci F. (2007), Pricing behavior and the response of hours to productivity shocks, Journal of Money, Credit and Banking, 39, 7, p.1587-1611.

Mc Callum J. (1987), Unemployment in Canada and the US, The Canadian Journal of Economics, 20, 4, p.802-822.

Pagan A. (1984), Econometric issues in the analysis of regressions with generated regressors, International Economic Review, 25, 1, p. 221-247.

Pagan A. (1986), Two stage and related estimators and their applications, Review of Economic Studies, 53, p. 517-538.

Annexe : Le cas des variables basées sur le prédicteur (PGR)

Pagan (1984) considère (section 2 – p.222-232) les modèles (« modèle 1 ») où seuls les PGR sont introduits dans l'équation (1) :

$$(1)\ y = \delta z^* + e$$

$$(2)\ z = z^* + \eta = W\alpha + \eta$$

Ce système peut être estimé soit par une estimation jointe (maximum de vraisemblance) ou par une procédure en deux étapes (MCO appliqués à (2) pour obtenir $\hat{\alpha}$, puis construire $\hat{z} = W\hat{\alpha}$, et substituer $\hat{z}$ pour z* dans (1), et MCO appliqués alors à (1) pour obtenir $\hat{\delta}$).

Le théorème 2 (p.224-225) implique qu'une procédure en deux étapes est efficiente asymptotiquement pour l'estimation du paramètre δ (les deux estimateurs, maximum de vraisemblance et à deux étapes ont alors la même distribution limite si l'estimation de α converge vers sa vraie valeur, ce qui est assurée par les propriétés des MCO appliqué à l'équation (2)).

Le théorème 3 (p.226-227) montre que l'estimation de la variance (des écarts-types) de $\hat{\delta}$ donnée par une estimation par les MCO appliquée aux deux étapes est généralement inconsistante (elle est généralement moins élevée que la variance qui serait retirée d'une estimation par le maximum de vraisemblance). Les t-statistiques sont alors généralement plus grands que les t-statistiques corrects. Il conviendrait alors d'appliquer la technique des variables instrumentales (EVI) car elle produit alors des écart-types consistants[5].

En revanche, si l'hypothèse δ=0 est la seule hypothèse à tester qui nous intéresse, l'estimateur MCO de la variance de $\hat{\delta}$ est consistante et les t-valeurs asymptotiques sont valides. La régression MCO de y sur $\hat{z}$ est donc efficiente et suffisante. Pour toute autre hypothèse, il est nécessaire d'utiliser la technique des variables instrumentales ou des doubles moindres carrés (moindres carrés à deux étapes) pour obtenir des estimations consistantes de la variance de $\hat{\delta}$ ((1) devient alors $y = \delta z + e + \delta(z^*-z)$). [6]

[5] Voir Dufour et Jasiak (2000) pour un traitement récent du problème posé par les PGR et notamment l'utilisation de la technique des variables instrumentales.

[6] Redding S., and Venables A. J., 2004, dans une équation de salaire introduisent des mesures d'accés au marché prédites par une équation de commerce. Puisque les valeurs prédites de régresseurs obtenus d'une première équation sont introduites dans l'équation centrale, les erreurs cette équation incluent les erreurs de la première équation. La présence de régresseurs estimés signifie que, comme dans les DMC (MC à deux étapes), les écart-types des MCO ne sont pas valides. Les auteurs utilisent alors les techniques de bootstrap (Efron et Tibshirani, 1993) pour obtenir des écart-types qui prennent en compte explicitement la présence des régresseurs estimés (chaque réplication bootstrap estime l'équation de première étape, génère les valeurs prédites des régresseurs, et estime l'équation de deuxième étape).