

DE L'INTÉRÊT DE LA RÉGRESSION MULTINIVEAU EN POLITIQUE COMPARÉE

Damien BOL

L'utilisation de la régression en science politique francophone

La régression est la technique d'analyse de données la plus répandue en science politique. Il s'agit d'une technique dite quantitative qui, à partir de données récoltées systématiquement comme des réponses à un sondage ou des résultats électoraux, mesure l'association entre une ou plusieurs variables dites « explicatives » et une variable dite « expliquée ».

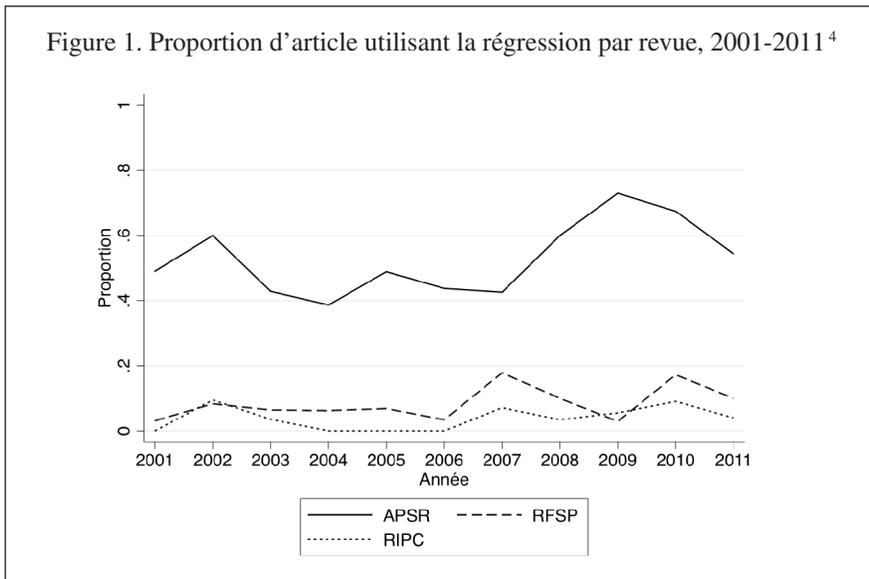
La régression possède de nombreux avantages. Premièrement, en donnant la possibilité de contrôler l'effet perturbateur de plusieurs autres variables explicatives, l'effet de chaque variable explicative sur la variable expliquée peut être interprété de manière causale. Deuxièmement, grâce à l'utilisation de p-valeurs, elle permet de générer de l'inférence, à savoir des conclusions qui s'appliquent au-delà des cas observés. Troisièmement, la régression est naturellement adaptée à un raisonnement de type hypothético-déductif qui est de plus en plus privilégié en science politique¹.

Ces dernières années, la régression s'est imposée comme la technique de référence dans les revues internationales les plus influentes en science politique. Par exemple, entre 2001 et 2011, environ 50 % des articles publiés dans l'*American Political Science Review* (APSR)² l'ont mise à profit, avec un pic à 73 % en 2007 (voir Figure 1).

1. Pour une initiation à la régression, aussi parfois appelée modélisation statistique, voir BRESSOUX P., 2008, *La modélisation statistique appliquée aux sciences sociales*, De Boeck Université, Bruxelles et PETRY F., GELINEAU F., 2009, *Guide pratique d'introduction à la régression en sciences sociales*, Presses de l'Université Laval, Québec.

2. L'APSR constitue certainement la revue de référence en science politique. Elle se veut généraliste et possède le facteur d'impact le plus élevé selon les standards développés par *Web of Knowledge*. En outre, l'APSR est ouverte à toutes les méthodes, tant qualitatives que quantitatives, et à la théorie politique.

A contrario, la régression reste relativement peu exploitée dans la science politique francophone. Par exemple, la proportion d'articles l'ayant utilisée lors de ces dix dernières années, que ce soit dans la Revue française de science politique (RFSP) ou dans la Revue Internationale de Politique comparée (RIPC)³, est marginale. Cette proportion n'a en effet jamais dépassé les 20 %. Récemment, une légère augmentation semble toutefois se dessiner. Dans la RFSP, la proportion d'articles utilisant la régression est ainsi passée de 3 % en 2001 à 9 % en 2011. En ce qui concerne la RIPC, il n'y a plus eu, depuis 2006, une seule année sans qu'au moins un article ne l'utilise (voir Figure 1).



Cette sous-utilisation de la régression dans les articles publiés dans la RIPC est étonnante. Le premier fondement de la revue est en effet de favoriser les analyses comparées⁵. Or, la régression possède plusieurs propriétés qui la rendent particulièrement appropriée à ce type d'analyses. Dans un article de 1997, Bruno Cautrès en faisait déjà état : étant insensible aux différences d'effectifs, la technique permet aux chercheurs de comparer les résultats d'analyses de sondages réalisés dans plusieurs pays. Ainsi, il préconisait de réaliser les mêmes analyses en parallèle dans chacun des pays étudiés et de comparer les résultats obtenus, en particulier les *odd ratios*

3. La RFSP est la revue principale de science politique en français. Selon les standards établis par l'Aeres, elle fait partie du premier groupe en termes d'influence sur la discipline. Étant généraliste et ouverte à tous les types de méthodes, elle peut être considérée comme le pendant francophone de l'APSR. La RIPC se situe quant à elle dans le second groupe de revues francophones les plus influentes de la discipline selon l'Aeres.

4. Les données proviennent d'une recension réalisée par l'auteur. La définition ayant servi à l'identification des articles utilisant la régression est celle développée dans l'article.

5. Voir la description de la RIPC, accessible via <http://www.ripc.eu>.

qui constituent des mesures normalisées et facilement interprétables des associations entre différentes variables⁶.

Bien qu'utile, cette démarche n'exploite toutefois pas l'ensemble du potentiel de la régression pour l'analyse comparée. Pour l'essentiel, elle ne permet que de mettre en lumière les différences qui existent entre les pays étudiés sans pour autant les expliquer. Le soin revient alors au chercheur de trouver une interprétation plausible qui ne peut hélas pas être directement confrontée aux données.

En intégrant directement la composante comparée dans l'analyse, la régression, dite « multiniveau », permet de résoudre ce problème. Cet article en présente les deux types les plus répandus en sciences sociales : la régression à effets fixes et celle à effets aléatoires. Ensuite, nous démontrons que ces sous-techniques permettent de répondre à deux enjeux fondamentaux de la politique comparée, à savoir la quête de l'universalité et le refus des explications culturalistes.

Fondamentaux de la régression

La régression est une technique d'analyse de données qui permet d'évaluer l'effet d'une ou de plusieurs variables explicatives (X_1, X_2, \dots, X_n) sur une variable expliquée (Y). Ses applications en science politique sont multiples. Elle peut par exemple servir à analyser les réponses à un sondage ou des résultats électoraux.

Afin d'illustrer notre propos, nous utilisons tout au long de cet article deux exemples hypothétiques : l'analyse de l'effet de la classe sociale d'un individu (X_1) sur sa propension à voter pour le parti socialiste en Belgique (Y) et l'effet de la couverture médiatique d'un candidat à une élection présidentielle française (X_1) sur son score obtenu (Y). Pour ce faire, nous avons généré deux bases de données fictives. La première analyse reprend les réponses à un sondage effectué sur un échantillon représentatif de la population belge de 100 personnes, comprenant leur vote lors de la dernière élection et leur classe sociale ; la seconde consiste dans les résultats électoraux de dix candidats à l'élection présidentielle française et leur couverture médiatique durant la campagne (en heures d'antenne)⁷. Afin de faciliter la compréhension du lecteur, les analyses peuvent être répétées en

6. CAUTRES B., « Modèles log linéaires et analyse comparative de données d'enquête », *Revue Internationale de Politique Comparée*, vol. 4, n° 1, 1997, p. 84.

7. Dans la pratique, les bases de données utilisées pour réaliser des régressions doivent contenir plus de cas : au moins 500 pour des analyses de données de sondage et au moins 30 pour des analyses de résultats électoraux.

téléchargeant les deux bases de données fictives et une syntaxe Stata sur le site internet de l'auteur (<http://www.damienbol.eu>).

Lorsque la question de recherche a été posée et que les données ont été récoltées, la relation hypothétique entre les variables d'intérêt doit être pensée théoriquement et un modèle doit être construit. La classe sociale a-t-elle un effet sur la propension à voter pour le parti socialiste pour chaque individu, ou est-ce que les jeunes électeurs échappent à cette règle ? Est-ce que la relation entre la couverture médiatique d'un candidat et son score à l'élection présidentielle est telle que, plus sa couverture est grande, plus son score est élevé ? Ou existe-t-il un effet de « plafond » au-delà duquel le candidat est déjà suffisamment connu de la population et que sa couverture médiatique n'a plus d'effet sur son score ?

Le modèle théorique le plus simple de relation entre des variables d'intérêt est le modèle linéaire. Il prend la forme suivante :

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon$$

Dans les exemples présentés ci-dessus, ce modèle correspond à l'hypothèse selon laquelle l'appartenance à la classe ouvrière ou la couverture médiatique d'un candidat (X_1) a un effet linéaire sur la propension à voter pour le parti socialiste ou sur le score de ce candidat (Y). D'autres relations entre ces variables d'intérêt auraient toutefois pu être postulées et d'autres modèles auraient pu être construits. Par exemple, une variable reportant l'âge de l'individu aurait pu être ajoutée au premier modèle pour rendre compte de l'hypothèse selon laquelle la classe ouvrière n'a d'effet sur la propension à voter pour le parti socialiste que pour les électeurs plus âgés. L'âge et la classe sociale auraient alors été multipliés pour rendre compte de leur interaction dans l'explication de la variable expliquée⁸.

De la même façon, le logarithme de la variable couverture médiatique aurait pu être ajouté au deuxième modèle pour rapporter l'effet de plafond mentionné ci-dessus. En principe, tout opérateur mathématique peut être utilisé. Il appartient au chercheur d'être original dans sa modélisation théorique. Dans un souci de clarté, cet article s'en tiendra aux modèles linéaires. L'ensemble des arguments développés ci-dessous s'applique toutefois à tout autre modèle.

Lorsque le modèle est construit, le chercheur peut estimer ses paramètres (β_0 et β_1) en le confrontant aux données empiriques et en appliquant une

8. Sur comment modéliser des effets d'interactions, voir BRAMBOR T., CLARK W., GOLDBER M., « Understanding Interaction Models: Improving Empirical Analyses », *Political Analysis*, vol. 40, n° 1, 2006, p. 63-82.

procédure mathématique consacrée à l'aide d'un logiciel statistique⁹. C'est cette opération qui porte le nom de régression. Le coefficient β_1 donnera alors une mesure de combien, en moyenne, Y augmente lorsque X_1 augmente d'une unité¹⁰. β_0 , la constante, prendra quant à elle la valeur de Y quand X_1 est égal à 0, ce qui dans nos exemples correspond à la propension à voter pour le parti socialiste pour les individus ne faisant pas partie de la classe ouvrière ou le score obtenu par un candidat à l'élection présidentielle qui ne bénéficie d'aucune couverture médiatique.

Le dernier élément de l'équation, ϵ , représente l'erreur standard qui est aléatoire et varie pour chaque cas. Son but est d'absorber la part de variabilité de Y qui n'est pas expliquée par les variables du modèle.

Tableau 1. Régressions simples

	Vote PS	Vote PS	Score candidat	Score candidat
Être ouvrier	4.88** (2.55)	1.03 (0.77)		
Être membre d'un syndicat		21.83** (17.04)		
Couverture médiatique			1.47** (0.05)	1.52** (0.02)
Faire partie du gouvernement sortant				-2.10** (0.24)
Constante	0.14** (0.06)	0.05** (0.03)	-0.27 (0.48)	-0.19 (0.15)
N	100	100	10	10

Note : Les entrées sont des *odd ratios* estimés sur base d'un modèle log-linéaire par le calcul du maximum de vraisemblance (deux premières colonnes) ou des coefficients estimés sur base d'un modèle linéaire par le calcul des moindres carrés. L'erreur standard est entre parenthèses. * p-valeur > 0.05 ; ** p-valeur > 0.01.

9. Les procédures d'estimations les plus fréquentes sont celles des moindres carrés et du maximum de vraisemblance. Les arguments développés dans l'article restent identiques dans les deux cas. Les logiciels les plus fréquemment utilisés en sciences sociales sont SPSS, Stata et R.

10. Lorsque la variable expliquée est binaire, comme c'est le cas pour le vote pour le parti socialiste, le modèle devra subir une transformation avant de pouvoir être estimé. En particulier, son logarithme népérien devra être calculé, le transformant en modèle log-linéaire. Après estimation, l'exponentiel de β_1 pourra être interprété en termes d'odd ratio ou en d'autres mots, de combien augmentent les chances de voter pour le parti socialiste lorsque X_1 augmente d'une unité. Voir CAUTRES B., *op. cit.*, 1997, p. 77-91. Les arguments de cet article s'appliquent également à ces modèles log-linéaires.

La Tableau 1 fait apparaître les résultats obtenus en estimant ces modèles linéaires à une variable explicative sur nos deux bases de données fictives (respectivement, la première et la troisième colonne). D'une part, nous observons que le fait d'être ouvrier augmente par un peu moins de cinq la propension d'un individu à voter pour le parti socialiste en Belgique par quatre par rapport aux autres électeurs¹¹. D'autre part, il apparaît que le score d'un candidat à l'élection présidentielle française augmente de 1,47 points lorsque sa couverture médiatique augmente d'une heure. Ces deux effets sont statistiquement significatifs à un niveau de p-valeur < 0,01.

Le modèle linéaire peut contenir plus d'une variable explicative. Les effets de ces dernières sur la variable expliquée s'additionnent alors de la façon suivante :

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n + \varepsilon$$

Ce type de modèle permet de s'assurer que les coefficients mesurent bien l'effet net de chaque variable explicative sur la variable expliquée. En effet, lorsque les variables explicatives sont associées entre elles (et avec la variable expliquée), la régression peut mener à des estimations erronées.

Dans notre premier exemple, le chercheur peut ajouter une variable rapportant le fait que le répondant est membre d'un syndicat. La syndicalisation est en effet plus répandue parmi les ouvriers que dans le reste de la population. Il se pourrait alors très bien que ce soit cette syndicalisation (X_2) et non l'appartenance à la classe sociale ouvrière (X_1) qui explique la plus grande propension à voter pour le parti socialiste (Y). Ainsi, lorsque l'appartenance à un syndicat est ajoutée au modèle, l'effet de la classe sociale disparaît. Dans la deuxième colonne du Tableau 1, il apparaît que les ouvriers ont seulement 3 % de chances de plus que les autres de voter pour le parti socialiste en Belgique (comme le laisse entendre la p-valeur, cet effet n'est pas statistiquement significatif). Par ailleurs, les membres d'un syndicat ont environ 20 chances de plus que les autres de voter pour ce parti (significatif à un niveau de p-valeur < 0,01). L'association entre la classe sociale et le vote pour le parti socialiste trouvée dans la première colonne du Tableau 1 n'est en fait qu'un sous-produit de l'association entre cette variable expliquée et une autre variable explicative.

De la même manière, l'effet de la couverture médiatique (X_1) sur le score d'un candidat à la présidentielle (Y) est susceptible d'augmenter lorsque le fait que le candidat est ou non issu de l'opposition (X_2) est également pris en compte dans le modèle. Les membres du gouvernement sortant sont en effet

11. Sur comment interpréter les *odd ratios*, voir CAUTRES B., *op. cit.*, 1997.

plus susceptibles que les autres candidats de disposer d'une large couverture médiatique. Toutefois, ils seront peut-être aussi plus susceptibles que les autres d'être punis par des électeurs mécontents de l'état de l'économie dans le pays. Il serait donc judicieux d'ajouter cette variable supplémentaire. Comme reporté dans la quatrième colonne du Tableau 1, l'effet d'une unité de couverture médiatique supplémentaire sur le score d'un candidat à l'élection présidentielle française est en réalité de 1,52 %. Aussi, il apparaît que le fait d'avoir fait partie du gouvernement sortant diminue le score d'un candidat de 2,10 %. Ces deux effets sont statistiquement significatifs à un niveau de p-valeur $< 0,01$.

Dans la pratique, le chercheur ajoute souvent de nombreuses variables explicatives dites de « contrôle » afin d'empêcher ce genre de sur- et de sous-estimation. C'est cette possibilité de contrôler par un grand nombre de variables explicatives et donc d'explications alternatives à la variable expliquée qui différencie la régression d'autres techniques d'analyse de données comme le coefficient de corrélation, le chi carré, ou le v de Cramer. Elle permet d'isoler l'effet d'une variable explicative sur une variable expliquée et, dès lors, de pouvoir interpréter cet effet en terme de cause à effet et pas seulement en termes d'association entre deux variables.

Deux objectifs de la politique comparée

À son lancement en 1994, les éditeurs de la *RIPC* ont voulu faire de cette revue une plateforme privilégiée pour les études comparées en français. Selon eux, la revue constituait, et constitue d'ailleurs toujours, une formidable promesse pour le développement de la discipline. Le but avoué est de favoriser l'analyse comparée de plusieurs pays. Dès le premier numéro, deux pères fondateurs de la *RIPC*, à savoir Daniel-Louis Seiler et André-Paul Frogner, ont mis en lumière deux objectifs plus précis : la quête de l'universalité et le refus des explications de type culturaliste.

D'une part, Seiler soulignait l'importance de comparer différents pays afin d'amener la science politique vers un savoir plus généralisé¹². Selon lui, les conclusions tirées par le chercheur doivent s'affranchir du contexte spécifique des données analysées et constituer de réelles explications universellement valides. Sur le modèle des sciences exactes, il souligne l'importance de travailler à l'élaboration de « lois », qui même si elles ne

12. SEILER D.-L., « Science politique, comparaison et universaux », *Revue Internationale de Politique Comparée*, vol. 1, n° 1, 1994, p. 91-111.

sont pas déterministes, permettent de mieux comprendre l'être humain en tant qu'animal politique.

Dans nos exemples, la proposition de Seiler reviendrait par exemple à trouver l'effet moyen par lequel l'appartenance à la classe ouvrière augmente la propension à voter pour le parti socialiste dans l'ensemble des pays européens et pas seulement en Belgique ou celui par lequel la couverture médiatique augmente le score des candidats aux élections présidentielles au-delà du cas français.

D'autre part, toujours dans le premier numéro de la *RIPC*, Frognier¹³, rappelant ainsi les prescriptions de Przeworski et Teune¹⁴, mettait en lumière l'impératif de ne pas se résoudre, lorsque plusieurs pays sont comparés, aux explications dites culturalistes. Ces dernières consistent à attribuer des différences observées dans deux ou plusieurs pays aux différences culturelles qui les séparent. Elles constituent, selon Frognier, davantage un constat d'ignorance qu'une réelle explication. Il est en effet trop facile pour un chercheur d'utiliser ce genre d'argument « fourre-tout » lorsqu'il est incapable d'expliquer des tendances divergentes. Face à ce refus, la politique comparée vise à investiguer les effets sur la variable expliquée à la fois des facteurs individuels, dits « micro », et des facteurs institutionnels ou contextuels, dits « macro », qui se situent à un niveau supérieur à l'unité d'analyse initiale.

Dans notre premier exemple, il pourrait être intéressant de comparer l'effet de l'appartenance à la classe ouvrière sur la propension à voter pour le parti socialiste en Belgique et en Allemagne. Dans le cas où cet effet serait différent, une explication pourrait reposer sur la différence d'offre politique à gauche dans les deux pays. Si en Belgique les socialistes constituent la seule option viable de ce côté du spectre électoral¹⁵, il n'en est pas de même en Allemagne où la classe ouvrière pourrait aussi se laisser convaincre par un autre parti de gauche comme *Die Linke*. Dans le même ordre d'idées, il pourrait être intéressant de comparer l'effet de la couverture médiatique sur le score des candidats aux présidentielles en France et aux États-Unis. Il se peut que l'effet soit plus important de l'autre côté de l'Atlantique où il n'y a aucun contrôle sur les dépenses de campagne. Ainsi, 100 individus issus d'un échantillon représentatif de la population allemande et dix candidats

13. FROGNIER A.-P. , « Logique(s ?) de la politique comparée », *Revue Internationale de Politique Comparée*, vol. 1, n° 1, 1994, p. 70-74.

14. PRZEWORSKI A., TEUNE H., *The Logic of Comparative Social Inquiry*, Wiley-Interscience, New York, 1970.

15. Du moins, jusqu'à l'élection de mai 2014.

à l'élection présidentielle américaine ont été ajoutés aux deux bases de données fictives¹⁶.

Dans cet article, nous entendons démontrer que les régressions multiniveau, et en particulier la régression à effets fixes et celle à effets aléatoires, sont particulièrement appropriées pour rencontrer ces deux objectifs fondamentaux de la politique comparée. Les présenter nous a semblé d'autant plus important qu'ils ont été très peu utilisés dans la science politique francophone jusqu'à présent. Les articles publiés dans la RIPC et la RFSP ces dix dernières années n'en comptent que trois¹⁷.

La régression multiniveau

La régression multiniveau repose sur les mêmes principes que ceux de la régression « simple » présentée ci-dessus. La différence se situe dans sa capacité à gérer des bases de données comprenant des observations issues de plusieurs pays et des variables explicatives se situant à deux niveaux d'analyse. Il s'agit d'un côté du niveau inférieur j (micro), comme les réponses des individus à un sondage ou les caractéristiques propres des candidats à une élection, et de l'autre du niveau supérieur i (macro), reprenant les caractéristiques du pays dans lequel ces électeurs ou candidats évoluent. Le niveau i constitue donc le groupe dans lequel le niveau j s'insère.

Mathématiquement, un modèle linéaire multiniveau prend la forme suivante :

$$Y_j = \beta_0 + \beta_{1j}X_{1j} + \beta_{2j}X_{2j} + \dots + \beta_{nj}X_{nj} + \beta_{1i}X_{1i} + \beta_{2i}X_{2i} + \dots + \beta_{ki}X_{ki} + \epsilon_j$$

Ce modèle est très similaire à celui de la régression linéaire simple. Il postule que la variable expliquée, dans nos exemples il s'agit toujours du vote pour le parti socialiste et du score d'un candidat à l'élection présidentielle (Y_j), est fonction de plusieurs variables explicatives. La différence se situe dans le fait que les variables explicatives peuvent se situer sur les

16. Dans la pratique, les bases de données utilisées pour réaliser des régressions multiniveau doivent contenir plus de cas au niveau supérieur : au moins dix pays pour les régressions à effets fixes et au moins 30 pays pour les régressions à effets aléatoires.

17. JADOT A., « (Ne pas) être un électeur européen. Une analyse multiniveaux des déterminants individuels et contextuels de l'abstention en 1999 », in *Revue Internationale de Politique Comparée*, vol. 9, n° 1, 2002, p. 31-45 ; LAVER M., BENOÎT K., « Le changement des systèmes partisans et la transformation des espaces politiques », *Revue Internationale de Politique Comparée*, vol. 14, n° 2, 2007, p. 303-324 et MAGNI BERTON R., « Pourquoi les partis gouvernementaux perdent-ils les élections intermédiaires ? Enquête Eurobaromètre 2004 et élections européennes », *Revue française de science politique*, vol. 58, n° 4, 2008, p. 643-656.

deux niveaux (X_j et X_i)¹⁸. Par exemple, l'appartenance à la classe ouvrière du répondant et la couverture médiatique du candidat sont des variables relevant du niveau inférieur (X_{ij}), alors que l'offre politique à gauche et le caractère strict du contrôle des dépenses de campagne dans le pays sont des variables relevant du niveau supérieur (X_{ii}).

Cependant, estimer ce modèle multiniveau avec une régression linéaire simple risquerait de violer une des exigences fondamentales de toute régression, à savoir l'indépendance des cas observés. Cette indépendance est généralement assurée dans les bases de données construites sur base de réponses à un sondage en sélectionnant aléatoirement un échantillon de répondants dans une population bien définie, comme par exemple, les citoyens belges. Chacun d'entre eux aura ainsi les mêmes chances d'être interrogé et les cas seront indépendants les uns des autres. En ce qui concerne les candidats à l'élection présidentielle française, la question ne se pose même pas puisque la base de données reprend le score de tous les candidats. Ils ont donc autant de chances les uns que les autres (à savoir 100 % des chances) d'être sélectionnés pour participer à l'analyse.

Cette indépendance des cas observés est remise en question dès que plusieurs pays sont comparés. Même si les cas sont indépendants à l'intérieur de chaque pays, ils ne le sont plus au niveau global de la base de données. Les répondants au sondage et les candidats sont regroupés en deux : respectivement, les répondants belges et allemands ou les candidats français et américains.

De plus, un autre problème à l'estimation de ce genre de modèle multiniveau avec une régression linéaire simple se situe dans l'hétérogénéité des cas observés. Dès que le chercheur se met à comparer plusieurs pays, il court le risque de se heurter à des phénomènes politiques fort différents. Si on considère qu'il existe de nombreuses différences entre les individus d'un même pays, ces différences sont certainement encore plus fortes entre des individus de pays différents. Ainsi, il est raisonnable de penser qu'un électeur belge a plus de caractéristiques communes avec un autre électeur belge qu'avec un électeur allemand. De la même manière, un candidat français partage sans aucun doute plus de similarités avec un de ses homologues français qu'avec un candidat américain. La non-prise en

18. Dans le cas où la variable explicative se situe au niveau supérieur, la régression multiniveau n'est pas requise. Dans pareilles situations, le chercheur peut se contenter d'agrèger les variables du niveau inférieur afin qu'elles coïncident avec celles du niveau supérieur et de réaliser une régression simple à un seul niveau. Par exemple, pour étudier l'effet du degré d'éducation (niveau inférieur) sur les soulèvements populaires dans les pays arabes (niveau supérieur), il suffit de modéliser ces soulèvements en fonction de la moyenne nationale du niveau d'éducation.

compte de ces différences peut mener le chercheur à estimer des effets imprécis, voire erronés.

La régression multiniveau permet de corriger ces deux problèmes. Dans cet article, nous en présentons deux types : la régression à effets fixes et la régression à effets aléatoires. La correction qu'ils amènent étant différente, chacun d'entre eux permet de rencontrer un des deux objectifs de la politique comparée présentés ci-dessus, à savoir la quête de l'universalité et le refus des explications culturalistes.

Les modèles à effets fixes : viser l'universalité

Les modèles à effets fixes résolvent le problème d'indépendance des cas présentés ci-dessus en ajoutant une variable explicative par pays (X_i). Dans notre premier exemple sur le vote pour le parti socialiste en Belgique et en Allemagne¹⁹, il prend donc la forme suivante :

$$Y_j = \beta_{0j} + \beta_{1j}X_{1j} + \beta_{2j}X_{2j} + \dots + \beta_{nj}X_{nj} + \beta_{1i}X_{Belgique\ i} + \beta_{2i}X_{Allemagne\ i} + \epsilon_j$$

Par rapport au modèle linéaire simple, une variable supplémentaire par pays est ajoutée ($X_{Belgique\ i}$ et $X_{Allemagne\ i}$)²⁰. Lors de l'estimation, les paramètres, dits « fixes », associés à ces variables (β_{1i} et β_{2i}) prendront alors la moyenne de la variable expliquée dans le pays et permettent ainsi de contrôler, en même temps, l'effet perturbateur de l'ensemble des facteurs se situant à ce niveau. Par exemple, ils permettent de contrôler les caractéristiques propres du système politique comme le mode de scrutin, ou celles des partis socialistes belges et allemands comme leur capacité à avoir gardé un fort ancrage local ou non.

Ainsi, la régression à effets fixes répond également au problème d'hétérogénéité des cas observés. Elle assure que l'effet des variables explicatives se situant au niveau inférieur ne soit pas biaisé par la non-prise en considération de facteurs se situant à un niveau supérieur, à savoir le pays. Elle permet dès lors de rencontrer l'objectif que Seiler attribue à la politique comparée : découvrir des effets qui soient universels.

19. La logique est la même concernant notre deuxième exemple sur le score des candidats aux élections présidentielles en France et aux États-Unis. Il suffit d'ajouter des variables supplémentaires par pays.

20. En pratique, il n'est pas nécessaire d'ajouter une variable explicative pour chaque pays. Il suffit d'ajouter une variable explicative pour chaque pays moins un. Le paramètre associé au dernier pays est repris par la constante (β_0).

Le Tableau 2 révèle les résultats de la régression à effets fixes dans nos deux exemples fictifs (respectivement, la première et la troisième colonne). Il apparaît que l'appartenance à la classe ouvrière augmente un peu moins que quatre fois la probabilité d'un individu de voter pour le parti socialiste (statistiquement significatif à un niveau de p-valeur < 0,01). Comme la base de données contient les réponses d'un échantillon représentatif de citoyens belges et allemands et qu'une variable par pays a été ajoutée au modèle (même si elles n'apparaissent pas directement dans le tableau), ce résultat est interprétable au-delà des deux pays étudiés. On peut raisonnablement penser que cet effet de la variable explicative sur la propension à voter pour le parti socialiste est généralisable à l'ensemble de la population de pays étudiés, en l'occurrence les pays européens.

De la même manière, le Tableau 2 laisse apparaître qu'une augmentation d'une heure de couverture médiatique augmente le score d'un candidat à l'élection présidentielle de 1,94 points. Ici aussi, l'utilisation d'une régression à effets fixes sur deux pays (France et États-Unis) permet de contrôler l'ensemble des facteurs de niveau supérieur qui diffèrent entre ces deux cas (par exemple, le fait qu'il y ait bien plus de conditions à remplir pour être candidat à l'élection présidentielle américaine ou que les médias publics y soient quasiment inexistantes). Ainsi, le résultat peut être également généralisé à l'ensemble des démocraties établies possédant un régime présidentiel.

Tableau 2. Régressions multi-niveau

	Vote PS (effets fixes)	Vote PS (effets aléatoires)	Score candidat (effets fixes)	Score candidat (effets aléatoires)
Être ouvrier (inférieur)	3.91** (1.45)	3.97** (1.48)		
Offre politique à gauche (supérieur)		0.74** (0.26)		
Couverture médiatique (inférieur)			1.94** (0.10)	1.94** (0.10)
Financements illimités (supérieur)				4.51** (1.59)
N	200	200	20	20

Note : Les entrées sont des odd ratios estimés sur base d'un modèle log-linéaire par le calcul du maximum de vraisemblance (deux premières colonnes) ou des coefficients estimés sur base d'un modèle linéaire par le calcul des moindres carrés. L'erreur standard est entre parenthèses. * p-valeur > 0.05 ; ** p-valeur > 0.01.

Les modèles à effets aléatoires : refuser les explications culturalistes²¹

La régression à effets fixes permet de générer une estimation non biaisée de l'effet d'une variable explicative se situant au niveau inférieur sur une variable expliquée se situant également au niveau inférieur généralisable au-delà des pays analysés. Toutefois, elle ne permet pas de mesurer l'effet d'une variable explicative spécifique se situant au niveau supérieur. C'est précisément ce que la régression à effets aléatoires permet de faire, tout en corrigeant le problème de non-indépendance des cas observés et de leur hétérogénéité.

La régression à effets aléatoires prend la forme suivante :

$$Y_j = \beta_{0j} + \beta_{1j}X_{1j} + \beta_{2j}X_{2j} + \dots + \beta_{nj}X_{nj} + \beta_{1i}X_{1i} + \beta_{2i}X_{2i} + \dots + \beta_{ki}X_{ki} + \epsilon_{i Pays} + \epsilon_j$$

Selon ce modèle, la variable expliquée à un niveau inférieur (Y_j) est fonction de variables explicatives situées à un niveau inférieur (X_j) et d'autres variables explicatives situées à un niveau supérieur (X_i). Pour corriger les problèmes liés au caractère multiniveau de la base de données, l'erreur standard ϵ est divisée en deux. Pour rappel, l'erreur standard est un terme aléatoire qui varie pour chaque cas observé et dont le but est d'absorber, pour chacun d'entre eux, la part de variabilité de Y_j qui n'est pas expliquée par le modèle.

Comme pour chacun des modèles présentés ci-dessus, le modèle associé à la régression à effets aléatoires prévoit une erreur standard au niveau inférieur (ϵ_j). Toutefois, à ce premier terme d'erreur, il en ajoute un deuxième au niveau supérieur (ϵ_i). Ainsi, le problème de non-indépendance des cas observés et de leur hétérogénéité est corrigé.

L'avantage de la régression à effets aléatoires est qu'elle permet d'estimer l'effet sur des variables explicatives se situant au niveau supérieur. En ce sens, elle donne la possibilité au chercheur de rencontrer l'objectif rappelé par Frogner (voir ci-dessus) d'expliquer les différences entre pays autrement qu'en faisant référence à leurs différences culturelles. Pour reprendre notre deuxième exemple fictif, en comparant plusieurs régimes présidentiels tels que les États-Unis et la France, le score des candidats à l'élection présidentielle peut être estimé en fonction de facteurs plus spécifiques tels que la couverture médiatique (niveau inférieur) et la loi sur le financement des campagnes électorales (niveau supérieur). L'analyse va au-delà du constat de l'existence d'une simple différence entre ces pays.

21. Par modèle aléatoire, nous entendons les modèles à interceptes aléatoires. Il existe en effet une autre classe de modèles dits à pentes aléatoires, moins utilisés en science politique. Nous ne les présentons pas dans cet article. Le lecteur intéressé pourra consulter STEENBERGEN M., JONES B., « Modeling Multilevel Data Structures », *American Journal of Political Science*, vol. 46, n° 1, 2002, p. 218-237.

Ainsi, comme reporté dans le Tableau 2 (spécifiquement, dans la colonne 4), la non-limitation des dépenses électorales (comme c'est le cas aux États-Unis) s'avère avoir un effet positif moyen de 4,51 % sur le score du candidat. En effet, ce système a tendance à favoriser les deux candidats principaux qui disposent de plus d'argent et qui se partagent donc la quasi-totalité des voix exprimées lors d'un scrutin. Cet effet est statistiquement significatif à un niveau de p-valeur $< 0,01$. En outre, il est aussi contrôlé, à un niveau inférieur, par la couverture médiatique du candidat lors de la campagne.

De la même manière, dans notre premier exemple (voir la colonne 2 du Tableau 2), l'existence d'un autre parti à la gauche du parti socialiste diminue la probabilité d'un individu de voter pour ce parti de 26 %, même en contrôlant par l'appartenance de cet individu à la classe ouvrière. Cet effet est aussi significatif à un niveau de p-valeur $< 0,01$.

Par ailleurs, il est important de noter que, contrairement aux modèles à effets fixes, les modèles à effets aléatoires ne corrigent pas l'éventuel biais de sélection des unités de niveau supérieur, en l'occurrence les pays. Pour que les résultats soient généralisables à l'ensemble de la population de pays étudiés, ils doivent être sélectionnés aléatoirement. Une des solutions utilisées dans la littérature scientifique pour résoudre ce problème est de sélectionner l'entièreté des observations de cette population macro, comme par exemple l'ensemble des pays européens ou les démocraties établies disposant d'un régime présidentiel.

Conclusion : vers une bonne utilisation de la régression multiniveau

Dans cet article, nous nous sommes efforcés de présenter, de manière simple et adaptée aux chercheurs non familiers avec les méthodes quantitatives, la régression multiniveau. Celle-ci étant peu répandue dans les principales revues de notre discipline, nous estimons qu'il est important de mettre en lumière ce qu'elle peut apporter à la science politique en général et à la politique comparée en particulier. À cet égard, nous démontrons en quoi elle permet de répondre à deux objectifs fondamentaux de cette sous-discipline, à savoir la quête de l'universalité et le refus d'explications de type culturaliste.

La régression multiniveau n'est toutefois pas exempte de défauts. Comme toute méthode quantitative, elle impose au chercheur à simplifier les concepts utilisés. Ce problème se fait d'autant plus sentir lorsque le nombre de pays comparés est grand. Le chercheur doit alors s'efforcer de trouver des concepts qui soient transposables dans divers contextes nationaux sans

pour autant perdre de vue le sens qui leur est propre dans ce contexte. De plus, chaque pays ayant un système politique au fonctionnement spécifique, la comparaison à grande échelle mène fatalement à une perte de complexité. Toutefois, nous pensons que ces critiques n'enlèvent en rien l'intérêt de ce type de méthode. Pour preuve, bien que ces critiques soient connues depuis longtemps²², elle continue à être utilisée, et l'est même de plus en plus ces dernières années.

22. DOGEI M., « L'analyse quantitative en science politique : us et abus », *Revue Internationale de Politique Comparée*, vol. 1, n° 1, 1994, p. 19-36.