Séminaire «Qu'est-ce que *Faire preuve* ? » CURAPP
5 Mai 2006

Henry Rouanet (CRIP5, Université René Descartes)
& Frédéric Lebaron (Université d'Amiens)

# La preuve statistique : Examen critique de la régression

1. Econométrie et régression
2. Pour un usage raisonnable de la régression
   2.1 Régression dans un contexte non-expérimental ; hyper-expérimentalisme. Prédiction et explication. Inférence statistique sur des données d'observation. Fit-&-test technique.
   2.2 Spécification des variables. Effet de structure. Quasi-colinéarité. « Effets vrais » des variables « toutes choses égales par ailleurs ». Illusions et pollution.

*Références*

Le Roux B. & Rouanet H. (2004), *Geometric Data Analysis*. (Avant-propos de Patrick Suppes, Stanford), Dordrecht : Kluwer. Chapitre 8 « Inductive Data Analysis » ; chapitre 9 « Case study : The French political space ».
Rouanet H., Lebaron F., Le Hay V., Ackermann W., Le Roux B. (2002), Régression et Analyse géométrique des données : réflexions et suggestions, *Mathématiques et Sciences Humaines*, p. 13-45.

Le texte qui suit (mis en forme en décembre 2006) constitue une partie d'un chapitre en cours de rédaction sur le thème : *Analyse des données et régression : espace social et sociologie des variables.* Ce texte a bénéficié des remarques de Julien Duval, que nous remercions vivement.

Ayant pris en compte l'ensemble des agents efficients (individus et, à travers eux, institutions) et l'ensemble des propriétés – ou des atouts – qui sont au principe de l'efficacité de leur action, on peut attendre de l'analyse des correspondances, qui, *ainsi utilisée, n'a rien de la méthode purement descriptive que veulent y voir ceux qui l'opposent à l'analyse de régression,* qu'elle porte au jour la structure des oppositions, ou, ce qui revient au même, la structure de la distribution des pouvoirs et des intérêts spécifiques qui détermine, et *explique*, les stratégies des agents, et par là l'histoire des principales interventions qui ont conduit à l'élaboration et à la mise en application de la loi sur l'aide à la construction

P. Bourdieu, *Les structures sociales de l'économie* (p.128)

Une diminution de cinq élèves par classe conduirait à une réduction de 45% des inégalités entre ZEP et non ZEP dans le primaire.(les Journaux[1])

## *Introduction*

Dans l'analyse statistique des données sociologiques, les méthodes d'analyse des données (AC et ACM) utilisées en France se sont trouvées, à partir des années 1980, concurrencées par les techniques de régression, linéaire puis logistique. Alors que l'Analyse Géométrique des Données est en affinité avec la représentation spatiale (au sens propre) de l'espace social, la régression renvoie à une ''sociologie des variables'' qui vise à établir les effets de facteurs sur une variable dépendante. Comme le rappelle Boudon (1967), l'usage de la régression en sciences sociales remonte au moins à Yule (1895, 1899). Dans les années 1950, la régression linéaire faisait partie des méthodes statistiques standard en sociologie[2], aussi bien qu'en psychométrie et bien sûr en économétrie.

Entre les sociologues dans la tradition de Bourdieu et les sociologues ''quantitativistes'', on a souvent assisté à un dialogue de sourds. On a aussi avancé l'idée de ''complémentarité''; mais quelle complémentarité ? S'il s'agit de concéder à l'Analyse des Données la phase descriptive et exploratoire, en réservant à la régression la phase explicative, cette sorte de complémentarité n'est pas acceptable. Selon nous, il n'est pas possible d'opposer des méthodes statistiques qui seraient, en quelque sorte par essence, ''exploratoires et descriptives'' à d'autres qui seraient ''explicatives'' et seules capables d'apporter la ''preuve statistique'' des conclusions avancées. A notre sens, plutôt qu'une complémentarité, on doit chercher une *synthèse* entre analyse des données et régression, visant à *intégrer* la régression dans l'analyse géométrique, et à *donner un sens* à la sociologie des variables, *replacée* dans le cadre de l'espace social. En vue de cet objectif, il nous faut au préalable procéder à un examen de la régression.

Notre plan sera le suivant :

Nous évoquerons d'abord la tradition de la régression en économétrie (§1).

Nous procéderons à un examen critique de la régression (§2).

Nous serons alors à pied d'œuvre pour aborder la discussion sur la sociologie des variables et l'espace social.

[1] *Le Monde,* Mardi 21 Février 2006, p. 23.

[2] Voir par exemple H.M. Blalock (1960), *Social Statistics*, McGrawhill; et R. Boudon (1967), *L'analyse mathématique des faits sociaux*, Plon.

## *1. Économétrie et régression*

Compte tenu du statut dominant de l'économétrie dans les sciences sociales ‘‘quantitatives’’, c'est vers l'économétrie que nous nous tournerons pour présenter la régression.

### 1.1. *Des phénomènes économiques aux modèles économétriques*

L'article fondateur de Frisch, en 1933, plaçait l'économétrie à la jonction de la théorie économique, de la statistique et des mathématiques[3]. Dans le traité classique d'Edmond Malinvaud (1981)[4] on lit (p.9): « Entendue dans un sens large, l'économétrie englobe toute application des mathématiques ou des méthodes statistiques à l'étude des phénomènes économiques… Dans le sens étroit… l'économétrie a pour objet propre la détermination empirique des lois économiques ». De fait, l'économétrie ne se réduit pas à une étude statistique ‘‘neutre’’ des phénomènes économiques tels que le produit intérieur brut, le chômage, l'inflation ou la masse monétaire ; mais elle renvoie toujours à un cadre théorique économique plus ou moins explicite[5]. Un *modèle économétrique*, par exemple macro-économique, permettant les prévisions et les variantes de politique économique, est la représentation simplifiée d'un objet d'étude, liée à un corps d'hypothèses théoriques.
Un constituant essentiel d'un modèle économétrique est la *visée prédictive.* Prédiction au sens temporel d'abord (on dit aussi prévision). Ainsi sur des séries temporelles, le modèle doit permettre, sur la base des valeurs connues au temps *t*, de prédire les valeurs des variables au temps *t+1*. Prédiction au sens large ensuite : à partir de ce qu'on connaît, on cherche à se prononcer sur ce qu'on voudrait savoir. Soit dans le langage technique de la *régression* : connaissant les valeurs observées d'un ensemble de *variables prédictrices* (dites aussi indépendantes ou exogènes), on estime la *variable à prédire* (dite aussi dépendante ou endogène). La régression est l'outil de base de l'économètre, avec son modèle-cadre : spécification des variables (indépendantes, dépendante) et fonction de lien (linéaire, logit…) exprimant les relations entre les variables.

*Modèles macro-économétriques*
Dans les débuts, l'économétrie était peu distincte de la modélisation des ‘‘grandes relations macro-économiques’’ : équation de consommation (la consommation nationale croît avec le revenu et décroît avec les prix), équation de prix (les prix augmentent avec le coût unitaire), relation de Phillips (le salaire moyen baisse quand le taux de chômage augmente) ; avec les systèmes d'équations simultanées, dont il s'agissait d'estimer les coefficients. Les données de base des modèles macro-économétriques sont les agrégats fournis par la comptabilité nationale et la statistique publique : PIB, importations, etc. ; elles sont souvent annuelles, parfois trimestrielles voire mensuelles.

[3] On trouvera dans l'ouvrage de A. Pirotte (2004), *L'économétrie : des origines aux développements récents*, Paris, éd. CNRS, une histoire et un panorama très documentés de l'économétrie. La thèse de M. Armatte (1995), *Histoire du Modèle Linéaire: formes et usages en statistique et en économétrie jusque en 1945,* et l'article d'Armatte (2005), La notion de modèle dans les sciences sociales, *Mathématiques & Sciences Humaines*, 172, 91-123 , contiennent des analyses détaillées des rapports entre régression et économétrie .
[4] Malinvaud (1981), *Méthodes statistiques de l'économétrie*, Dunod.
[5] L'économétrie est souvent associée à la théorie néo-classique, issue de Walras et Pareto, avec un versant ‘‘keynésien’’ et un versant ‘‘libéral’’ ; l'un et l'autre pouvant être mobilisés dans la construction d'un modèle. Pour une analyse des oppositions dans le champ de la science économique, cf. Lebaron (2000).

*Dossiers exemplaires*

*Dossier Malinvaud* (1981, p.19). Pour illustrer la régression, Malinvaud prend dans son traité une série temporelle dans laquelle les ‘‘individus’’ sont les années successives, de 1949 à 1966. La variable dépendante est la variable *importations* ; les variables prédictrices sont le PIB, les *stocks*, la *consommation* ; auxquelles on adjoint (pour prendre en compte le ‘‘décollage’’ après 1960), une *variable temporelle* valant 0 jusqu'à 1960, 1 en 1961, 2 en 1962, 3 en 1963, etc.
Le dossier Malinvaud a pour nous valeur de référence, car en subordonnant la méthodologie statistique à la problématique de recherche, il met l'accent sur les choix majeurs, à commencer par la *spécification du modèle*, qui doit comporter toutes les variables pertinentes et seulement ces variables. La littérature économétrique contient nombre de dossiers exemplaires ; on en trouvera dans les *Éléments d'économétrie* du site de Michel Volle[6], qui énonce clairement les trois phases essentielles de la modélisation 1) Schéma théorique ; 2) Spécification du modèle ; 3) Estimation des paramètres.

## 1.2 *Économétrie sans modèle aléatoire*

Ainsi que le rappelle Armatte (1995, 2005), il existait dans les années 1920 et 1930, notamment en France, une statistique ‘‘mathématique’’ (entendons mathématiquement élaborée) mais ‘‘non-probabiliste’’, autrement dit dont les modèles-cadres ne faisaient pas intervenir l'aléatoire. Dans cette longue tradition, on peut mentionner les *bunch maps* de Ragnar Frisch, évoqués par Malinvaud (1981) dans le premier chapitre de son traité, précisément intitulé «Econométrie sans modèle aléatoire»[7] ; les thèses des économistes du National Bureau of Economic Research (NBER), qui entendaient fonder l'analyse économique sur l'étude descriptive des cycles des affaires (cf. Pirotte, 2004, p. 31). Last but not least, Maurice Allais (lui aussi ‘‘prix Nobel d'économie’’) récuse énergiquement (Allais, 1983) toute ‘‘intervention du hasard’’ dans la théorie économique.
Ne pas mettre la charrue avant les bœufs : lorsque Benzécri déclarait « Le modèle doit suivre les données, non l'inverse ! », il situait clairement l'Analyse des Données dans cette tradition statistique ‘‘sans modèle aléatoire’’[8]. Dans les années 1970-1980, dans les techniques de l'administration statistique française, notamment en matière de dépouillement d'enquêtes par questionnaires, une place importante avait été acquise par l'Analyse des Données[9], très présente à l'INSEE dans les enseignements théoriques aussi bien que dans les recueils de données commentées, dans des publications régulières comme les *Données sociales*. Le statut (sinon l'usage effectif) des méthodes géométriques dans les études économiques va ensuite en déclinant dans les années 1980[10]. A l'heure actuelle, dans la hiérarchie des méthodes ‘‘reconnues’’, l'Analyse des Données est cantonnée aux études exploratoires, ou à des types de données spécifiques.

---

[6] www.volle.com/rapports/econometrie.htm

[7] S'agissant de la régression linéaire, il n'y a nul besoin de supposer un processus aléatoire pour évaluer les coefficients de régression et le *R* multiple; il suffit de procéder à un ajustement par les moindres carrés, ainsi que le fait Malinvaud. Ironie du sort : le titre courant du chapitre (reproduit en haut des pages) est «Econométrie *sans modèle* (sic)»!

[8] Dans les *Cahiers de l'Analyse des Données*, et dans le volume F. & J.P.Benzécri, *Pratique de l'analyse des données en économie*, Paris, Dunod, 1986, tome 5, Benzécri a appliqué ses méthodes aux données économiques et financières.

[9] Le manuel de Michel Volle (1982), *Analyse des Données*, Economica, marque sans doute la ‘‘pointe avancée’’ de la pénétration de l'Analyse des Données à l'INSEE .

[10] Voir l'Histoire de l'AD. Ainsi la représentation des nuages d'individus, inconnue des travaux anglo-saxons, a vu son usage progressivement amenuisé dans les études françaises.

### 1.3. *Le modèle aléatoire en économétrie*

Le modèle aléatoire, aujourd'hui dominant, «ne s'est pas imposé dès le début de l'économétrie puisqu'il n'est apparu clairement formulé que dans le texte de Haavelmo[11]» nous dit Malinvaud (1981, p.3). Voir aussi Pirotte, (p. 47), Armatte (1995). À propos de ce manifeste de 144 pages, on a parlé de ''révolution probabiliste'': « No tool developed in the theory of statistics has any meaning - except perhaps for descriptive purposes (*sic*) - without being referred to some stochastic scheme.» proclame Haavelmo. L'économétrie aurait-elle attendu 1944 pour découvrir la probabilité?[12] En réalité, le paradigme dans lequel a basculé l'économétrie a été celui d'une idéologie statistique particulière alors en train de conquérir le pouvoir académique : *l'école fréquentiste radicale* de Neyman-Pearson. Or cette idéologie, qu'on peut qualifier de ''sample-minded'' (hors du *random sample* point de salut !), ne connaît, en fait de probabilités, que le modèle de départ censé représenter le processus aléatoire générateur des données[13]. Quoi qu'il en soit, l'alpha et l'oméga de la régression seront désormais respectivement le modèle aléatoire (avec son cortège d''‘assumptions'' : normalité, homoscédasticité, etc. ) et les tests d'existence des effets (et accessoirement les limites de confiance des coefficients).

*La position de Malinvaud.* En lisant en continuité les analyses du chapitre 1$^{er}$ (sans modèle aléatoire) puis celles du chapitre 6 (avec modèle aléatoire), on reconstitue *de facto* la primauté de la phase descriptive. «En économie, il est clair que les irrégularités des données ne résultent généralement pas de véritables tirages au sort. Aussi verrais-je grand intérêt à l'établissement d'une statistique subjectiviste reposant sur le principe de Bayes.»[14] En somme, tout en faisant allégeance au modèle aléatoire, Malinvaud prend ses distances.

### 1.4. *De la macro-économétrie à la micro-économétrie*

À partir des années 1970, la modélisation macro-économique fait l'objet de critiques (cf. Pirotte, chap.5). Un ensemble d'innovations (anticipations rationnelles, modèles VAR, théorie des cycles réels…) déplacent les débats existants. L'audience croissante de la théorie néo-classique s'accompagne d'une référence aux fondements micro-économiques de la macroéconomie, à partir des comportements d'acteurs rationnels abordés selon l''‘individualisme méthodologique''. L'intérêt se déplace alors des séries macroéconomiques (fondées sur les agrégats de la comptabilité nationale) vers les comportements des entreprises, des ménages ou des personnes, par exemple en matière de consommation et épargne, d'offre et de demande de travail ou encore en matière éducative. Les données statistiques pertinentes sont celles d'enquêtes sur les ménages, les entreprises, ou des données administratives individuelles… À partir de ces données, on cherche à estimer les coefficients (par exemple les élasticités) de théories micro-économiques. Le passage d'une économétrie centrée sur les grandes relations macro-économiques à une économétrie préoccupée des comportements des acteurs est reflété par l'opposition entre les ''prix Nobel '' de 1969, R.Frisch, ou de 1980,

---

[11] Haavelmo (1944), the probability approach in econometrics, *Econometrica*, vol. 12 (supplement).

[12] Des ouvrages comme le *Treatise on Probabilities* de Keynes (1921) conduiraient à en douter, même si l'on sait que Keynes (avec d'autres) séparait soigneusement probabilité et économie: cf . Armatte (1995). Quoiqu'il en soit, un changement de paradigme aussi soudain est sans pareil dans les autres domaines concernés par la statistique. Ce phénomène mériterait une histoire sociale, qui examinerait entre autres s'il a eu des effets sur les pratiques en matière de politique économique.

[13] Elle ne permet pas de calculer les *probabilités des hypothèses* à partir des données, ou d'interpréter un intervalle de confiance en termes de probabilité sur le paramètre. V. Rouanet & al (1998), *New Ways in Statistical methodology: from significnce tests to Bayesian inference*, Berne, Peter Lang.

[14] La révolution bayésienne – authentiquement probabiliste, puisque la probabilité recouvre son statut de mesure de l'incertitude – est de nos jours en train de s'accomplir. Voir Rouanet & al (1998).

L.Klein, et ceux de l'an 2000, J.Heckman et D.Mc Fadden ; ou encore, à l'INSEE par le contraste entre l'enseignement de E.Malinvaud dans les années 1960, et celui de C. Gouriéroux depuis les années 1980.

### 1.5. *Régressions pour variables ''qualitatives''*

Le passage de la macro à la micro-économétrie s'est accompagné de l'usage croissant des régressions spécialement conçues pour les variables catégorisées (dénommées ''qualitatives'') : log-linéaire, logit, etc. Riandey (1991) parle du modèle logit comme d'une méthode « pratiquée à l'INSEE depuis une bonne dizaine d'années »[15].

En régression linéaire, pour traiter ces variables (en particulier les variables dichotomiques), on employait (on emploie toujours) le codage en 0,1 (variables indicatrices, alias *dummy variables*). Ce codage entraîne des propriétés indésirables, comme de conduire à des fréquences prédites en dehors de l'intervalle [0,1][16]. Des modèles comme le modèle logit, qui modélise le logarithme du rapport des chances (LogOdds) sont très utiles dans des domaines où on travaille sur des petites fréquences, comme l'épidémiologie. Pour des fréquences éloignées de 0 et 1, les propriétés mathématiques de la fonction logistique entraînent que les fréquences prédites par le modèle logit sont proches de celles du modèle linéaire; proximité qui, il faut le dire, constitue une surprise pour maint utilisateur.
Il ne faut pas confondre avancée technique avec percée théorique; ce que laisseraient entendre les discours stéréotypés expliquant que les ''nouvelles régressions'' permettraient désormais de «séparer et quantifier les effets purs » des variables. Nous examinerons ces allégations au §2.

*Une fuite en avant technologique ?*

Lorsqu'on ouvre les actuels manuels d'économétrie[17], on constate que l'arsenal technologique s'est colossalement accru. Outre l'incontournable *General Linear Model*, nous avons maintenant les modèles logit, tobit pour variables "qualitatives". Nous trouvons aussi la panoplie impressionnante des *tests préalables* des *assumptions* : normalité, orthogonalité, suridentification, homoscédasticité (Goldfeld & Quandt, Breusch-Pagan, White, Glesjer), indépendance sérielle (Durbin-Watson, Ljung & Box), etc.[18] Plutôt qu'une logique scientifique, cette inflation technologique évoque une logique d'hôpital : le patient (nous voulons dire le modèle) est soumis à une batterie de tests ; si les résultats sont bons (i.e. non-significatifs), il est élargi - du moins jusqu'à plus ample informé[19].
Sans doute, les économètres dans leur pratique de recherche (voir les dossiers ci-après) «en prennent et en laissent ». Mais la surenchère technologique a un effet dissuasif auprès des

[15] Voir Gouriéroux (1989), *Econométrie des variables qualitiatives*, Economica ; Lollivier, Marpsat, Verger (1991), *L'économétrie et l'étude des comportements: modèles de régression qualitatifs*. La référence de base est D.R. Cox (1970), *Analysis of binary data*, London, Methuen.

[16] On a la même sorte de propriété indésirable avec le modèle normal classique de la régression linéaire, lorsqu'on l'applique à des variables intrinsèquement bornées, par exemple non-négatives.

[17] V. par exemple le manuel fort bien documenté de Claudio Araujo, Jean-François Brun & Jean-Louis Combes (2004) : *Econométrie*, Bréal.

[18] Les étudiants qui peuvent réciter la panoplie méritent certainement leur UV. Mais savent-ils mieux franchir le pont-aux-ânes de l'économétrie (dixit Malinvaud), à savoir distinguer les situations où la quasi-colinéarité est nuisible et celle où elle ne l'est pas? A propos de certains problèmes méthodologiques, on trouve des assertions déconcertantes. Ainsi pour les tests de signification, là où Malinvaud suggérait au moins d'adapter le seuil à la taille de l'échantillon, les auteurs du manuel précité signalent au passage: «L'usage consiste à ne pas changer le seuil; les grands échantillons entraînent donc plus facilement que les petits le rejet de H0. » Acceptation résignée d'une pollution jugée inéluctable ?

[19] O mânes de Popper, les économètres ont enfilé tes chausses !

chercheurs qui ne sont pas rompus à ces exercices d'école et ne vivent pas toujours bien leurs « insuffisances mathématiques». [20]

### 1.6. *La pratique économétrique : la "fit-&-test technique"*

En fait, dans la pratique économétrique, l'analyse statistique se réduit souvent à la variante expéditive de la doxa : on prend un ensemble plus ou moins large de variables, on fait tourner les programmes de régression et on commente les effets statistiquement significatifs. C'est l'examen de cette ''*fit-&-test technique*'' qui nous retiendra désormais.
Pour se convaincre que l'usage de la régression tel que nous le décrivons est une pratique très commune, il suffira d'évoquer quelques dossiers.

*Le dossier Cukierman*[21]
Le problème étudié est l'influence de l'indépendance de la banque centrale sur l'inflation dans les économies ex-socialistes (entre 1989 et 1998). Les ''individus'', au nombre de $n = 57$, sont des combinaisons pays-années. La variable dépendante est l'*inflation*, définie comme le taux de dépréciation de la monnaie $D= 1/(1+F)$ (où $F$ est l'indicateur d'inflation usuel). Cinq variables indépendantes sont considérées : 1) Indice d'*indépendance légale* de la banque centrale ; 2) la présence de *guerre* ou non (variable dichotomique) ; 3) Indice de *libéralisation de l'économie;* 4) Indice de *libéralisation des prix* ; 5) Indice d'*indépendance multiplicative* de la banque centrale (indice amendé obtenu en mettant à zéro l'indice d'indépendance légale tant que la libéralisation de l'économie n'a pas atteint un certain niveau). La conclusion est qu'une fois que le processus de libéralisation est bien enclenché, l'indépendance légale devient efficace pour réduire l'inflation; la guerre ayant par ailleurs un effet inflationniste (phénomène bien connu). Une version simplifiée du dossier Cukierman (retenant les seules variables 1, 2 et 5) nous servira d'illustration didactique: V. Encadré au §2.

*Dossier McFadden*
Dans son article sur le ''comportement de choix''[22], McFadden prend comme illustration les données d'une enquête (réalisée à Pittsburgh sur 140 personnes à la fin des années 1960) sur le choix du mode de déplacement pour faire ses courses. La variable dépendante est la variable dichotomique: voiture individuelle vs autre moyen de transport. Plusieurs modèles sont présentés. Le plus simple (modèle 1) retient 4 variables prédictrices : 1) temps de déplacement à la marche, 2) temps de parcours en voiture 3) coût du déplacement automobile, 4) rapport du nombre de voitures au nombre d'actifs dans le ménage. Le plus complexe (modèle 2) ajoute à ces 4 variables deux variables dichotomiques : la race (blanc *vs* non-blanc) et la profession (cols bleus *vs* cols blancs). Selon Mc Fadden, tous les coefficients trouvés « ont les signes attendus».[23]

---

[20] En réalité, les *mathématiques* ne sont pas en cause, mais leur adéquation à la problématique de recherche. Il ne faut pas se tromper de cible en dénonçant les ''abus des mathématiques''.

[21] Cukierman A., Miller G.P., Neyapti B. ((2000) , Central Bank reform, liberalization and inflation in transition economies; an international perspective, *Journal of Monetary Economics.*

[22] Mc Fadden (1973) Conditional logit analysis of qualitative choice behaviour, *Frontiers in econometrics*, Zarembka , Academic press, 105-142.

[23] Ce dossier illustre un phénomène déconcertant (devenu de nos jours monnaie courante): le fantastique décalage entre la sophistication de la théorie économique sous-jacente et la minceur des conclusions de l'analyse statistique censée étayer cette théorie; ce qui renvoie au problème épistémologique majeur de la validation d'une théorie individuelle à partir de données agrégées (cf. *Annexe* sur la théorie de l'action rationnelle). La perception de ce décalage n'apparaît guère dans les attendus de l'Académie des Sciences de Suède qui ont valu à McFadden l'attribution du ''prix Nobel '' «Quels sont les facteurs qui déterminent si une personne choisit de travailler ou non, et dans l'affirmative, combien d'heures? Comment les incitations

*Le dossier Fitoussi*
Jean-Paul Fitoussi & Coll *Réduction du Chômage : les Réussites en Europe*, rapport du Conseil d'analyse économique Paris, *la Documentation française.*
Les ''individus '' sont les 21 pays de l'OCDE. La variable dépendante est la variation du taux de chômage entre deux périodes; les variables indépendantes sont les variations de caractéristiques institutionnelles.

*Dossier Herpin-Verger*[24]
Les données sont issues de l'enquête Budget de famille de 1995 (n=11000 ménages). Elles portent sur les dépenses de consommation des ménages relatives à différents types de biens et services. Les auteurs étudient l'effet de variables telles que le revenu, la catégorie sociale, l'âge, sur les dépenses.

*Dossier Crepon-Desplatz (2001)*
Le fichier analysé est constitué par n=90000 entreprises pour lesquelles on dispose de diverses variables (effectif, part des non-qualifiés, rapport capital/travail… ) et d'un indicateur de baisse des charges sociales employeurs (laquelle induit une baisse du coût de travail peu qualifié). On étudie l'effet de cette baisse sur l'emploi, ainsi que sur la production et la productivité , en contrôlant diverses variables structurelles, telles que la structure de la concurrence, les indicateurs de performance financière.

### 1.7. *Vers l'absorption de la sociologie quantitative par les modèles économétriques ?*

Avec les travaux de micro-économétrie en matière de marché du travail, d'éducation ou de consommation, l'économétrie englobe désormais toutes les formes de comportements et de pratiques économiques et sociales. Selon Hendry (1995) (cité par Araujo & al) « L'économétrie devient l'approche scientifique visant à la compréhension des aspects économiques de la conduite humaine. » . On rejoint de la sorte les thèmes traditionnels de la sociologie quantitative (enquêtes par questionnaire, etc.), qui de son côté a importé les techniques de régression pour données catégorisées. La frontière entre économétrie et sociologie statistique devient très floue, avec la régression comme instrument commun de base. Avec l'usage en sciences sociales calqué sur l'économétrie, ''modèle économétrique'' en vient à désigner tout modèle de régression même sans lien avec une problématique économique.

Ce qui distingue l'approche économique de l'approche sociologique des mêmes phénomènes est donc moins désormais les techniques utilisées que les schémas explicatifs et partant les variables indépendantes retenues. C'est peut-être surtout la référence plus ou moins étroite au cadre théorique de l'*homo economicus* rationnel qui les distingue. Les économistes ont tendance à privilégier les variables du type « revenu » et « coût relatif » pour expliquer des

économiques influencent-elles les choix de formation, de métier et de lieu de résidence? Quels sont les effets de différents programmes de formation pour le marché du travail sur les revenus et l'emploi? …À partir de sa théorie économique sur les choix discrets, McFadden a développé de nouvelles méthodes statistiques qui ont eu une influence décisive sur la recherche empirique... McFadden n'a pas hésité à les utiliser lui-même (*sic*) dans des applications pratiques telles que la conception du réseau express régional de San Francisco ou les investissements dans les services téléphoniques et les résidences pour personnes âgées. »

[24] Nicolas Herpin & Daniel Verger (1999).Consommation et stratification sociale selon le profil d'emploi, *Economie et Statistique*, 324-325.

comportements[25]. Les sociologues quant à eux accordent une place primordiale aux variables indiquant des positions (y compris au sein de réseaux sociaux, dans le contexte de la nouvelle sociologie économique) et des dispositions (notamment culturelles)[26].

Mais socialement, le rapport de forces entre les disciplines est dissymétrique : l'économétrie est dominante, la sociologie dominée[27]; comme en témoigne l'étonnant « questionnaire » adressé par des économistes « à leurs collègues sociologues »[28]. Sachant que dans l'analyse économique, une fois identifiés les principaux facteurs de la croissance, on constate un résidu important, il est demandé aux sociologues « de découvrir les rôles qu'ont pu jouer dans leur mise en œuvre les divers facteurs sociologiques auxquels on peut penser. » A ce questionnaire, C. Baudelot devait apporter une réponse pleine d'esprit, renvoyant à Bourdieu & Darbel (Darras, 1966), pour un modèle typiquement économétrique, et offrant à la méditation des économètres une Analyse des Correspondances.

[25] Voir le cas exemplaire de l'économétrie du vote, discuté en détail dans les travaux de P.Lehingue sur l'analyse économique du comportement politique, Politix.

[26] Comme on sait, John Goldthorpe étudie les relations entre des variables explicatives telles que la position de classe (d'origine ou d'appartenance) et des variables à expliquer (niveau scolaire atteint, position professionnelle, opinions, pratiques culturelles, vote, etc.). Dans la mouvance de Goldhorpe, on évoquera les recherches de Goux & Maurin sur la démocratisation, celles de L.A. Vallet sur la mobilité sociale, ou encore celles de Coulangeon sur les déterminants sociaux des pratiques culturelles.

[27] Comment interpréter ce phénomène d'absorption ? Faute d'une histoire sociale des pratiques statistiques en sciences sociales, qui reste incomplète, on peut néanmoins risquer quelques hypothèses, comme celle selon laquelle la sociologie de Bourdieu n'est pas parvenue à faire reconnaître la méthodologie statistique qu'elle a mise en œuvre.

[28] J.J. Carré, Dubois P. & E. Malinvaud (1972).*La croissance française.* Le Seuil, Annexe 8, 6676-670. C. Baudelot (1988). Confiance dans l'avenir et vie réussie, *Mélanges économiques en l'honneur d'Edmond Malinvaud*, Economica.

***Annexe:*** *Théorie de l'action rationnelle et régression*

La théorie de l'action rationnelle ou RAT (*Rational Action theory*: Becker, R. Boudon…) inspire nombre de travaux de sciences sociales. Les théoriciens de l'action rationnelle dérivent à partir d'hypothèses portant sur les comportements individuels (individualisme méthodologique) certaines relations entre variables. Ce sont ces relations qui font l'objet de confrontation statistique. Par exemple, dans l'*Inégalité des chances* (Boudon), on postule que les familles décident chaque année du fait que les enfants poursuivent ou non leurs études, et que les familles modestes tendent à sous-évaluer le gain à retirer d'une année d'étude supplémentaire (et à en sur-évaluer le coût). Elles tendent ainsi à s'auto-éliminer progressivement. Ce processus conduit à une sous-représentation des catégories populaires dans l'enseignement supérieur, conforme aux données observées. La même démarche (calcul coût-bénéfice) est utilisée pour le choix d'avoir un enfant (Gary Becker), etc.
La remarque fondamentale de Pierre Bourdieu (1975)[29] est que l'accord des prédictions et des données au niveau agrégé ne constitue pas une validation du processus individuel hypothétique de choix rationnel, auquel Bourdieu préfère des anticipations différenciées fondées sur des *ethos* de classe différents.
En somme, Bourdieu considère que la théorie du choix rationnel pose de *bons problèmes* mais apporte de *mauvaises réponses*. Au lieu de se tourner vers l'ensemble des variables sociologiques pertinentes, elle réduit le choix à un arbitrage coût-bénéfice, en se bornant à certains facteurs économiques. Cette réduction intellectuelle se paye pour Bourdieu en capacité prédictrice et, plus encore, explicative.
De ce dernier point résulte la critique de l'usage de la régression, instrument privilégié de la RAT: un modèle incomplet et réducteur, qui ne permet au mieux qu'une prédiction imparfaite.

[29] P.Bourdieu, « Avenir de classe et causalité du probable », *Revue française de sociologie*, 1975.

## 2. ***Regard critique sur la régression***

La régression est une méthode puissante et utile. Mais il n'y a pas de voie royale en statistique : toute méthode comporte des difficultés[30] et requiert des précautions d'emploi. Notre propos dans ce qui suit n'est pas une vaine dénégation rituelle de ‘‘la portée et les limites’’ d'une méthode (la régression), mais de pénétrer dans la ‘‘boîte noire’’, au risque de mettre en cause l'idéologie dominante[31]… Pour un bon usage de la régression, les mises en garde n'ont pas manqué. Outre aux enseignements des économètres distingués, nous ferons appel à ceux du grand statisticien W. Tukey, partiellement reproduits dans F. Mosteller & J. Tukey (1977) ; et à ceux de Guttman (1977)[32].

Nous évoquerons d'abord la régression dans un contexte non-expérimental, avec schéma explicatif; nous commenterons la tentation de l'hyper-expérimentalisme ; nous reviendrons sur la spécification du modèle ; et nous commenterons l'usage des tests de signification.

Nous définirons ensuite les effets de structure en régression, et nous aborderons la difficulté centrale de la régression: la quasi-colinéarité, avec le dilemme exhaustivité-parcimonie et le problème de la grandeur des effets.

Pour conclure, nous donnerons notre réponse à la question : la régression permet-elle de dégager des "effets vrais, toutes choses égales par ailleurs", et d'apporter une "preuve statistique"?

Les *extraits du dossier Cukierman* (encadré ci-après) nous serviront d'exemple didactique, pour rappeler les concepts fondamentaux de la régression, et illustrer les discussions.

[30] Nous ne disons pas "pièges", la notion de piège statistique ne faisant pas partie de notre bagage intellectuel.
[31] Pour contester une idéologie, il ne faut pas trop compter sur les dominés, lesquels, même lorsque leur bonne connaissance des dossiers leur permet de contourner les conclusions artefactuelles, se retranchent volontiers pour la méthodologie dans l'attitude modeste du ‘‘praticien qui fait ce que tout le monde fait’’.
[32] Voir F. Mosteller & J. Tukey (1977). *Data Analysis and Regression*. Reading, Addison-Wesley; et Guttman (1977), What is not what in statistics, *The Statistician*, 26, 2, 81-107.

Extraits du *Dossier Cukierman (cf. § 1)*

Nous considérerons les 3 variables prédictrices réduites (i.e. centrées et d'écart-type unité) notés *x1* (indépendance légale de la Banque Centrale), *x2* (guerre) et *x5* (indépendance multiplicative de la BC); la variable à prédire réduite sera notée *y* (inflation). On a *n* = 57 "individus" (pays-périodes)

Toutes les régressions peuvent être obtenues à partir du *tableau des corrélations :*

| | *y* | *x1* | *x2* |
|---|---|---|---|
| *x1* | -.464 (S**) | [1] | |
| *x2* | +.425 (S**) | -.236 (NS) | [1] |
| *x5* | -.585 (S**) | +.931 (S**) | -.220 (NS) |

Les coefficients des *régressions simples* sont les coefficients de corrélation (les variables sont réduites) ; ils définissent les *effets globaux* de chacune des variables:

*y* sur *x1*: *y1*= - 0.464 *x1* ; avec $R^2$ = .215
*y* sur *x2*: *y2*= + 0.425 *x2*; avec $R^2$ = .181
*y* sur *x5*: *y5*= - 0.585 *x5* ; avec $R^2$ = .342

L'effet global est donc négatif (réduction de l'inflation) pour l'indépendance légale et pour l'indépendance multiplicative; il est positif (inflationniste) pour la Guerre

Dans les régressions multiples, les coefficients des variables réduites (*beta-weights*) définissent les *effets conditionnels*.

La *qualité de l'ajustement* d'une régression est définie par $R^2$ (carré du coefficient de corrélation multiple *R*), qui est la proportion de variance de la variable dépendante prise en compte ("expliquée") par les variables prédictrices [Pour les régressions simples, $R^2$ est le carré du coefficient de corrélation]. La valeur de $R^2$ est un indicateur de l'importance des *effets conjoints* des variables indépendantes. Quand on enrichit l'ensemble des variables, le $R^2$ ne peut que croître.

*Régression double* sur *x2* et *x5* :

*y25* = + 0.311 *x2* - 0.517 *x5*; avec $R^2$ = .435

[Effets conditionnels : + pour *x2* (S**); - pour *x5* (S**).]
Cette régression double est le résumé le plus compact des données.

*Régression triple sur (x1,x2, x5)*:

*y125*= + 0.690 *x1* + 0.335 *x2* - 1.154 *x*5; avec $R^2$ = .498

[Effets conditionnels : + pour *x1* (S*) ; + pour *x2* (S**); - pour *x5* (S**).]

*Interprétation des beta-weights*. Exemple: si $\beta 5$ est le coefficient de *x5* (variable réduite) dans une régression, lorsque la variable Indépendance multiplicative augmente d'un écart-type, les autres variables restant constantes, la variable Inflation s'accroît de $\beta 5$ fois l'écart-type. Ainsi dans la régression sur (*x2,x5*): lorsque l'indépendance multiplicative augmente d'un écart-type, la variable Guerre restant constante, l'inflation diminue de 0.517 écart-type. Dans la régression sur (*x1,x2,x5*) : lorsque l'indépendance multiplicative augmente d'un écart-type, l'indépendance légale et la variable Guerre restant l'une et l'autre constantes, l'inflation diminue de 1.157 écart-type.

N.B. Dans la régression triple, on note les signes opposés pour les deux variables *x1* et *x5,* qui toutes deux opérationnalisent l'indépendance de la banque centrale…

2.1. *Régression dans un contexte non-expérimental: prédiction vs explication*

La prédiction est l'essence de la régression. On peut faire de la régression un usage seulement *prédictif* ; dans ce cas la liste des variables prédictrices n'est pas limitative, et les liaisons entre ces variables sont peu gênantes[33]. Un usage plus ambitieux de la régression est l'usage *explicatif*. Dans un schéma explicatif, la variable à prédire est qualifiée de "variable à expliquer", et les variables prédictrices des "variables explicatives"; on voudrait connaître les "poids relatifs" des variables prédictrices dans le schéma explicatif. Notre discussion portera sur la régression dans un schéma explicatif.

La problématique prédictrice est *symétrique*. A partir de la température, on peut prédire la longueur d'une barre métallique, et régresser la longueur sur la température; mais on peut tout aussi bien, à l'inverse, à partir de la longueur, prédire la température (thermomètre), et régresser la température sur la longueur. La problématique explicative est *dissymétrique* : dans la théorie physique de la dilatation, la longueur est expliquée par la température, mais non l'inverse.

L'usage intempestif de la phraséologie "explicative" Ten régression oblige à rappeler le principe «Corrélation n'est pas causalité» - une constante de la méthodologie statistique. Le plus sage assurément serait de dissocier schéma explicatif (physique, économique, sociologique) et méthode statistique (régression en l'occurrence), en bannissant toute phraséologie explicative dans la phase d'analyse statistique[34]. Il n'y a pas de méthode statistique qui serait par essence "explicative"[35].

*Les trois commandements de Tukey*

Dans le contexte *expérimental*, les variables indépendantes sont des *facteurs* expérimentaux, dont on étudie les *effets* sur la variable dépendante à l'aide de méthodes classiques comme l'analyse de variance; techniquement, la régression n'est autre que l'analyse de variance pour des facteurs expérimentaux quantitatifs. Les succès de la régression en physique, aimait à rappeler Tukey, portent sur des modèles expérimentaux avec des variables indépendantes obéissant aux ''*trois commandements*'': 1) few in number ; 2) well-clarified ; 3) measured with small error (Mosteller & Tukey, 1977, p.321); auxquels on adjoindra le ''*quatrième commandement*'': terme d'erreur de faible variance (autrement dit $R^2$ pas trop petit)[36].

Pour les données d'observation, le modèle de régression est seulement un "modèle des données (*model of data*); les appellations "indépendantes" (pour "prédictrices") et "dépendante" (pour "à prédire"), sont métaphoriques[37]; on peut les garder, justement comme des rappels salutaires de l'allégeance de la régression (et de la sociologie des variables !) à l'égard de la méthodologie expérimentale[38] ; en particulier chaque fois qu'on considère des variables "sur lesquelles on pourrait agir", comme la flexibilité de l'emploi (dossier Fitoussi) ou l'indépendance de la banque centrale (dossier Cukierman).

[33] Ce qui importe pour la prédiction, c'est un coefficient $R^2$ aussi élevé que possible ; dans cette perspective, l'idée d' « effet vrai, toutes choses égales par ailleurs » n'est pas pertinente.

[34] « La statistique n'explique rien, mais elle fournit des éléments potentiels d'explication » écrivent Lebart et al. (1995), *Statistique exploratoire multidimensionnelle*, Dunod. On ne peut que souscrire à cette déclaration. J.P. Fénelon, qui avait le sens des formules, résumait ainsi le débat cornélien du chercheur : « Dire descriptif, c'est péjoratif ; dire explicatif, c'est abusif. »

[35] En psychométrie, la tendance serait plutôt de voir la régression comme un outil pragmatique de prédiction (pronostiquer le résultat à un examen à partir de résultats à des tests); et de donner un statut explicatif à l'analyse factorielle (dans le modèle unifactoriel, le facteur d'intelligence générale $g$ " explique " les résultats aux tests). La régression n'a pas le monopole de l'explication. On peut faire également - Goodman (1976), promoteur de la régression, n'hésite pas à le recommander - un usage exploratoire de la régression.

[36] Rappelons la boutade (attribuée à Tukey): «If you want to explain something, you must have something to explain.». Peut-on sérieusement considérer qu'un $R^2$ inférieur à .05 mérite d'être " expliqué"?

[37] L'absconse terminologie des économètres (exogène vs endogène) n'a pas été adopté en sciences sociales.

[38] Allégeance finement analysée par Passeron (1991, p. 129), *Le raisonnement sociologique*, Nathan.

*Hyper-expérimentalisme*

Les coefficients de régression sont des *statistiques descriptives*[39]: elles résument les données, au même titre que des moyennes ou des variances. Un résumé statistique n'est pas un système dynamique de forces conflictuelles. *L'hyper-expérimentalisme* est la sur-interprétation la plus grossière des résultats de la régression. Il consiste à s'imaginer que la valeur d'un coefficient révèlerait l'ampleur de l'effet qu'on est en droit d'attendre si, dans la réalité, on modifiait la valeur de la variable indépendante correspondante. Malheureusement, dans un système complexe, modifier la valeur d'une variable peut entraîner un *changement structurel* du modèle-cadre. Comme le souligne Haavelmo lui-même (1944, p. 26), une série statistique peut être ajustée par une relation simple ; puis une nouvelle série statistique, portant sur la même variable, encore par une relation simple…mais différente de la première! L'épigraphe de ce chapitre (effet à attendre de la diminution du nombre d'élèves par classe) est un exemple d'hyper-expérimentalisme[40].
Mosteller & Tukey (1977, p. 320) posent carrément la question: «We want to know what will actually happen when we change a variable»; et donnent à méditer la réponse du grand statisticien G.E.P Box (qui ne passe pas pour un extrémiste) : «The only way to find out what will happen when a complex system is disturbed is to disturb the system, not merely to observe it passively.»

*Spécification du modèle-cadre*

Le choix des variables du modèle-cadre est le cœur de la régression: un modèle *correctement spécifié* est celui où figurent toutes les *variables pertinentes* (indépendantes, dépendantes), et seulement ces variables.

*Erreurs de spécification*

Pour Malinvaud (1981 chap 2 §16, chap.4, p. 354), l'omission de variables exogènes importantes est une des causes les plus fréquentes d'erreurs ; il donne (p.127) un exemple célèbre dans lequel une équation consommation-revenu conduisit à des prévisions erronées, en dépit d'un $R^2$ très élevé. Dans Malinvaud (1995), il revient à la charge: «Les petits écarts-types des estimations, ou les hauts niveaux de signification, ne sont valables que si la spécification est correcte.» (On songe à Pascal: «L'omission d'un principe mène à l'erreur. »)
Mosteller & Tukey (p. 328) donnent d'autres exemples fameux, tels que celui de la précision du tir des bombardiers lors de la seconde guerre mondiale. Parmi une dizaine de variables retenues (altitude, type d'appareil…), la variable "présence/absence de la chasse adverse" fut trouvée avec un coefficient élevé, mais de signe contraire au bon sens: "chasse adverse présente, meilleure précision du tir". Investigation faite, on s'aperçut qu'on avait omis la variable météorologique " temps clair / brumeux", d'où l'explication présumée: par temps clair, la précision est meilleure, mais la chasse adverse tend à se manifester.
Pour autant, il ne suffit pas de prendre toutes sortes de variables plus ou moins plausibles pour éviter les erreurs de spécification (V. la discussion de l'*exhaustivité* plus loin). En sciences sociales, Guttman (1977) rejoint Malinvaud et Tukey en déclarant : «Nothing may be more practical for arriving at a simplified regression than a substantive theory for the structure of the entire covariance matrix ».

[39] Techniquement: une *statistique descriptive* ne dépend pas de la taille des données : cf. Rouanet & al (1997).
[40] La sur-interprétation s'applique à toutes les variantes de régression. Autre exemple lu aussi récemment dans la presse: « S'il était interdit de fumer dans les bars, restaurants et discothèques, en l'espace de six mois, le nombre de décès consécutifs à un infarctus diminuerait de 40%». Ah ! ces journalistes ! Ne soyons pas trop durs pour eux! Où vont-ils chercher tout ça, sinon chez les services de communication des "scientifiques" ? Les coefficients de régression ne se prêtent que trop bien à la logique des "parts de facteurs". « M. le Professeur, dites-le à vos auditeurs : dans les accidents de voiture quels sont les pourcentages de responsabilité du conducteur et de la route? Quelles sont dans l'intelligence les parts respectives de l'hérédité et du milieu? etc. » Logique journalistique, dont la pauvreté contraste singulièrement avec la subtilité des modèles causaux que développait naguère un Boudon (1967).…

*Tests de signification*

L'inférence statistique n'est pas réservée à la méthodologie expérimentale; mais cette dernière pèse lourdement en régression. Dans le contexte expérimental, la notion de *replication* (reprise de l'expérience "dans des conditions identiques") permet à la rigueur de se figurer les probabilités comme limites de fréquences (cadre dit "fréquentiste" de l'inférence statistique), dans la ligne du modèle aléatoire posé au départ. Avec les données d'observation, la seule situation où ce cadre d'inférence est opératoire est le sondage aléatoire: on extrait un échantillon au hasard, et on souhaite étendre les conclusions à la population. En dehors de cette situation, le cadre fréquentiste est le plus souvent forcé. Par exemple, l'"échantillon" est un ensemble fixé d'années (dossier Malinvaud) ; ou l'ensemble de tous les pays d'Europe; ou encore (dossier Cukierman) un ensemble de pays-périodes d'économie ex-socialiste.
Aujourd'hui les seuils observés (les *p-values*) sont fournis par l'ordinateur - avec le *star-system* "en prime" : *p* entre *.05* et *.01* simplement significatif, une étoile S* ; *p< .01* hautement significatif, deux étoiles: S**, etc. Dans la régression sur 3 variables du dossier Cukierman, les coefficients de *x2* et de *x5* sont hautement significatifs, celui de *x1* simplement significatif. Faudrait-il se priver des tests de signification?

Il ne faut pas s'en laisser imposer par la pensée *sample-minded* « Hors de l'échantillonnage aléatoire point de salut ! ». Il faut savoir que d'autres cadres d'inférence existent, qui eux aussi permettent de formaliser la notion d'"effet vrai", dont l'effet observé est une estimation, et partant d'apprécier si un effet observé notable peut ou non être regardé comme une simple coïncidence "due au hasard"[41]. Quel que soit le cadre d'inférence, plus un résultat est significatif, mieux on est à même d'affirmer que l'effet vrai va dans le sens de l'effet observé; alors qu'un résultat non-significatif veut dire que le sens de l'effet vrai est incertain. Dans le dossier Cukierman, pour le modèle-cadre à trois variables *(x1,x2,x5)*, les résultats des tests corroborent les conclusions descriptives: inflationniste pour la Guerre et anti-inflationniste pour l'indépendance multiplicative; mais aussi (quoique dans une moindre mesure, une seule étoile) inflationniste pour l'indépendance légale[42].

*Significance fallacy*

Certains chercheurs, renonçant à estimer les coefficients de régression, se contentent de la variante la plus expéditive de la *fit-&-test technique*. Entendu à un congrès : «Moi, je regarde les signes des coefficients, et j'interprète ceux qui sont significatifs.» Avec cette pratique, le risque de *significance fallacy* n'est jamais loin. Ce risque est l'effet pervers de la propriété mathématique suivante des tests de signification, en soi irrécusable, mais qu'il faut toujours garder à l'esprit : plus la taille de l'échantillon est élevée, plus pour une même valeur donnée (non-nulle) d'un coefficient de régression, celui-ci est significatif[43]. En conséquence, avec un petit échantillon, un effet descriptivement important peut être non-significatif ; avec un grand échantillon, un effet descriptivement négligeable peut être significatif. Si le dossier Cukierman portait sur un effectif inférieur à 20 (au lieu de $n = 57$), avec les mêmes valeurs des coefficients, aucun des trois effets ne serait significatif.

[41] Il faut bien sûr dans chaque situation bien spécifier l'"hypothèse du hasard", c'est-à-dire «l'idée d'une chose qui pourrait *tout aussi bien* avoir lieu que ne pas avoir lieu.» (Nous reprenons une formulation de J. Bouveresse (1993), *Robert Musil, le hasard, la moyenne et l'escargot de l'histoire*, Paris, Editions de l'Eclat). C'est ainsi que le cadre des *tests de permutation* consiste à situer le protocole observé parmi un ensemble de protocoles possibles engendrés; un coefficient significatif signale un protocole observé *atypique*. Quant au cadre *bayésien*, il permet, moyennant des suppositions convenables, d'interpréter les seuils observés en termes de probabilités d'hypothèses. Cf. Le Roux & Rouanet (2004, p. 299-300).

[42] Dans la régression plus complète à 5 variables finalement retenue par Cukierman & Coll dans leur article, le coefficient de *x1* est positif, mais il n'est pas significatif, ce qui (dans la logique de la *fit-&-test* technique) coupe court à la difficulté de l'opposition des effets des deux variables d'indépendance de la banque centrale..

[43] Techniquement, le seuil observé ( *p-value* ) n'est pas une statistique descriptive.

L'illusion de significativité (*significance fallacy*) consiste à confondre effet significatif et important d'une part ; effet non-significatif et négligeable d'autre part. Les manuels de statistique mettent en garde contre cette illusion, avec les deux *Warnings* classiques : "Statistical significance is not substantive significance" et "No evidence of effect is not proof of no effect". Les experts relèvent parfois des abus criants[44]. Rien n'y fait[45].

## 2.2. *Effets de structure et quasi-colinéarité*

Même en supposant une spécification correcte des variables, à ce point commence la difficulté centrale propre à la régression. Pour la discussion qui va suivre, il suffira d'examiner de plus près la régression linéaire, avec les notions d'*effets globaux, effets conditionnels,* et *qualité de l'ajustement* $R^2$ (cf. Encadré Cukierman).

### *Effets de structure*

Dans les données expérimentales, les facteurs contrôlés sont en général organisés selon un plan orthogonal, et l'effet conditionnel d'un facteur est égal à son effet global. Pour les données d'observation, la corrélation entre variables prédictrices, c'est-à-dire la non-orthogonalité, est la règle, et l'effet conditionnel d'une variable diffère en général de son effet global. Dans l'article (Rouanet et al., 2002), nous sommes partis de la notion d'*effet de structure classique*, et nous avons étendu cette notion à la régression.

*L'effet de structure classique* est un fait bien connu en économétrie et démographie. Il consiste à constater que lorsqu'une population est divisée en sous-populations, une variable peut évoluer dans des sens différents selon qu'on la considère globalement (donnée agrégées) ou à l'intérieur des sous-populations. En particulier, elle peut évoluer à l'intérieur des sous-populations à l'inverse du sens global : c'est le *cas paradoxal* de l'effet de structure, qui dans les années 1930 avait beaucoup frappé les esprits : Simiand, Halbwachs[46]… Sur le site INSEE, on trouve une définition de l'effet de structure qui se focalise sur le cas paradoxal : «Une grandeur peut évoluer dans un sens sur chaque sous-population et dans le sens contraire sur l'ensemble de la population. » Pour une discussion approfondie, voir Michel Lévy & al (1981), qui commentent l'exemple: « Le salaire moyen des salariés dépend à la fois du niveau des salaires et de la part respective des hauts et des bas salaires. Si le niveau des salaires restant fixe …la proportion du personnel qualifié augmente aux dépens du personnel non-qualifié, le salaire moyen augmente sans qu'aucun salarié ne ressente d'amélioration.» [47]. Dans la littérature anglo-saxonne, l'appellation *Simpson paradox* renvoie à un article (Simpson, 1951) qui redécouvrant le cas paradoxal, véhicule une acception limitative de l'effet de structure. Dans les textes sur l'effet de structure classique, on ne trouve pas, à notre connaissance, d'allusion à la régression (en tout cas, pas dans l'article de Simpson).

---

[44] On rejoint ici les réserves de Malinvaud (en commentaire du dossier Fitoussier mentionné plus haut): «Il est maladroit et même erroné de ne retenir des analyses économétriques que les résultats ''significatifs au seuil de 5%''… De fait, dans les années récentes certains ont souvent rejeté l'idée de réformes institutionnelles qui auraient vraisemblablement amélioré l'emploi, au motif que ces réformes n'auraient pas un tel effet puisque les estimations des économètres ne conduisaient qu'à des résultats ''non significatifs''. »

[45] *Un exemple de significance fallacy* (parmi des milliers). « In Northern Ireland, the *lack of a significant effect implies* (les italiques sont de nous) that the chances of ending in the lowest level of Q category are *the same* for sons of service class (I +II) and for routine nonmanual class (III) origins » (Ishida & al, 1995, *Amer. J. Sociology* Comment les referees, d'ordinaire si pointilleux, laissent-ils passer de telles sur-interprétations ? Gageons que dans les synthèses sur la mobilité sociale figure désormais, au titre de "fait bien établi", qu' en Irlande du Nord, les chances sont les mêmes, etc.

[46] Dans Halbwachs (1935*), La statistique en sociologie,* à propos des taux de mortalité rectifiés (pour éliminer l'effet de structure), on lit : «Si la composition par âge de la population en France était modifiée, rien ne prouve que les taux de mortalité par âge ne seraient pas plus élevés », suivi de la remarque : « Nous ne contestons pas que ces procédures de représentation schématique n'aient leur utilité. » . Ces commentaires très équilibrés nous paraissent toujours d'actualité, et transposables à la régression.

[47] Michel Lévy & al (1981), *Comprendre l'information économique et sociale* , Hatier. Site INSEE : Hwww.insee.fr/fr/nom_def_met/definitionsH. Voir aussi Michel Novi (1998) : *Pourcentages et tableaux statistiques*, Paris, PUF.

En vue de la généralisation à la régression, nous sommes partis de l'effet de structure classique, illustré par notre "paradoxe des lycées". Dans une ville avec deux lycées, on compare les réussites des filles et des garçons, d'une part globalement (données agrégées), d'autre part à l'intérieur des lycées. Les résultats des deux types de comparaisons peuvent être de mêmes sens ou non ; le cas paradoxal est celui où les sens sont inverses : meilleure réussite des filles pour un type de comparaison, meilleure réussite des garçons pour l'autre .
Reformulons maintenant la situation précédente en termes de régression : la variable Réussite est la variable dépendante, les deux variables Sexe et Lycées deux variables prédictrices. L'étude de l'effet de structure classique devient celle du rapport Effet conditionnel / Effet global lié à la variable Sexe. Cette idée conduit à notre notion générale d'effet de structure en régression.

Pour caractériser la notion d'effet de structure en régression, nous considérons, pour chaque variable prédictrice, le rapport Effet conditionnel / Effet global, ce qui conduit à distinguer trois situations, que nous représentons par la "rose des vents des effets " (Rouanet & al, 2002 p. 26)
1) *Atténuation :* l'effet conditionnel est de même sens que l'effet global, mais moins marqué que celui-ci ; à la limite on a *disparition* de l'effet (effet global mais pas d'effet conditionnel) ; la situation de *stabilité* (effet conditionnel égal à l'effet global) étant intermédiaire entre atténuation et accentuation.
2) *Accentuation :* l'effet conditionnel est de même sens que l'effet global, mais plus marqué; à la limite on a *émergence* de l'effet (pas d'effet global mais effet conditionnel).
3) *Renversement* : l'effet conditionnel est de sens inverse de l'effet global; cette situation généralise le cas paradoxal de l'effet de structure classique. En régression, la situation paradoxale du renversement se rencontre fréquemment, surtout lorsqu'on prend beaucoup de variables prédictrices.

Dans le dossier Cukierman, dans la régression sur *(x2, x5)*, l'effet conditionnel de la Guerre (*x2*) (+0.311) est moins marqué que l'effet global (+0.425); on a donc atténuation; pour l'indépendance multiplicative (*x5*), l'effet conditionnel (-0.517) est également moins marqué que l'effet global (- 0.585) : on a aussi atténuation. Dans la régression sur (*x1,x2,x5*), pour la Guerre (*x2*) on a encore atténuation (+0.335 vs +0.425) ; mais pour l'indépendance multiplicative (*x5*) on a maintenant accentuation (-1.154 vs –0.585) ; et pour la variable indépendance légale (*x1*) on a renversement (0.690 vs –0.464) [48]

*Quasi-colinéarité*

Les chercheurs qui travaillent dans une problématique explicative souhaiteraient dissocier les effets des variables indépendantes. Dans la controverse sur le QI, ils voudraient bien évaluer les "parts respectives" de l'hérédité et du milieu. Malheureusement, lorsque deux variables sont fortement corrélées il peut être très malaisé de dissocier leurs effets, en ce sens qu'une modification infime des données peut entraîner des modifications considérables des coefficients : c'est le phénomène de *quasi-colinéarité*, qui s'étend au cas de plusieurs variables (on dit alors souvent *multicolinéarité)* [49].

[48] *Effet de structure et interaction*. Il y a là deux concepts différents. Si étant donnés deux lycées, les filles réussissent mieux que les garçons dans l'un des lycées, moins bien dans l'autre, on parle souvent d'interaction entre les facteurs Lycée et Sexe (encore une appellation en général métaphorique pour les données d'observation). Mais deux variables non-corrélées peuvent se trouver en interaction vis-à-vis d'une variable dépendante *y* (paradigme expérimental de l'interaction) ; et à l'inverse, deux variables corrélées peuvent avoir des effets additifs sur *y*, donc sans interaction.

[49] Dans la littérature économétrique, la quasi-colinéarité est souvent introduite par le biais de l'inférence statistique : la quasi-colinéarité rend les estimations des coefficients très incertaines. L'étude à l'aide de la formalisation linéaire permet un diagnostic plus précis des méfaits de la quasi-colinéarité (texte en cours de rédaction).

La quasi-colinéarité se produit en particulier lorsque des variables qui opérationnalisent à peu près la même entité sont introduites simultanément dans la régression : les coefficients de régression peuvent prendre des valeurs extravagantes voire ininterprétables[50].

*Exemples de quasi-colinéarité*
*Dossier Cukierman.* Les variables *x1* et *x5*, qui constituent deux variantes de la variable Indépendance de la banque centrale, sont fortement corrélées : Corr (*x1,x5*) = 0.931. Dans la régression sur (*x2,x5*), le coefficient de *x5* est négatif et très net: –0.517. On pourrait (naïvement) s'imaginer qu'en adjoignant la variable *x1*, on allait trouver deux coefficients négatifs plus petits ; or dans la régression sur (*x1,x2,x5*), l'effet de *x5* s'accentue (coefficient encore plus négatif -1.154), tout en étant en quelque sorte "compensé " par un coefficient positif pour *x1* (+0.690).
*Dossier Malinvaud.* Le coefficient, et même le signe, de la variable PNB dépend de la présence dans la régression de la variable Consommation, ou de la variable temporelle, ou des deux.
*Dossier Fitoussi.* (p.48). Le coefficient de la variable coordination entre employeurs passe de 2.2 à –2.4 si dans le modèle on adjoint la variation du chômage.

*Le dilemme exhaustivité-parcimonie*
Le remède radical à la quasi-colinéarité existe : réduire le nombre de variables. Entendu dans un congrès : «Moi, quand j'ai deux variables corrélées à .80, je ne prends que l'une des deux.» Cette pratique ne va pas sans risque[51]. De fait, le chercheur se trouve souvent placé devant le dilemme : "*Exhaustivité"* (prendre un ensemble de variables pertinentes aussi complet que possible, au risque de coefficients ininterprétables) ou "*parcimonie"* (prendre peu de variables mais qui peut-être ne sont pas les bonnes).

*Comparer les grandeurs des effets ; paradoxe de Frédéric*

Faute d'évaluer des valeurs numériques précises, peut-on au moins affirmer qu'un effet est plus grand qu'un autre ?[52] La réponse peut être ambiguë. Dans la régression sur (*x1,x2,x5*) du dossier Cukierman, laquelle des deux variables, Indépendance légale (*x1*), ou Guerre (*x2*) est la plus inflationniste ? La comparaison des beta-weights (+0.690 vs +0.335) suggèrerait que c'est l'indépendance légale ; mais l'effet de l'indépendance légale est moins significatif (*p*= .012, S*) que celui de la Guerre (*p*= .001, S**), ce qui suggère le contraire. Cette discordance entre beta-weights et *p-values* renvoie à une propriété mathématique indiscutable ; nous l'avons baptisée familièrement *paradoxe de Frédéric*, paradoxe qui peut se produire dès qu'on a plus de deux variables, et qui se produit effectivement, comme on pourra le constater aisément[53].
Comparer les grandeurs des effets en régression : est-il question plus naturelle ? Cette question reste apparemment, encore à l'heure actuelle, largement ouverte[54].

[50] Fisher, dès 1938, écrivait à propos d'une régression sur 7 variables: «This does not mean that the formula is worthless, but that all the individual coefficients might be varied largely, and provided the other coefficients were suitably adjusted, the predictive value would not appreciably be impaired. The formula in seven variables is therefore, to a large extent arbitrary and one or more of the variables used must certainly be redundant.»
[51] Soient deux variables *x1* et *x2* corrélées à .80, et soit *y= 5/3 x1 - 4/3 x2* (variables réduites). La variable *y* est en liaison linéaire parfaite avec *x1* et *x2* (corrélation multiple *R* égale à 1). Si l'on garde seulement la variable *x1*, la corrélation descend à .60 (corrélation *r1* entre *y* et *x1*); mais si on garde seulement *x2* , la corrélation tombe à zéro (la corrélation *r2* entre *x2* et *y* est nulle). Voir Guttman (1977).
[52] Joie de l'environnementaliste lorsqu'il peut annoncer: « Quand dans la régression on introduit outre le facteur Race le facteur Environnement, l'effet du facteur Race vient toujours après celui de l'Environnement.»
[53] Dans la régression sur les cinq variables (*x1,x2,x3,x4,x5)* retenue par Cukierman (cf. dossier au §1), si on ordonne les beta-weights du plus grand au plus petit (en valeur absolue), les variables (*x1,x2)* se rangent selon l'ordre [1] > [2] ; mais si on range les *p-values* de la plus significative à la moins significative, on trouve [2] > [1] ; on a donc encore le paradoxe.
[54] Dans des textes en cours de rédaction, nous proposons nos suggestions concernant le dilemme exhaustivité – parcimonie et le problème de la grandeur des effets.

## 2.3. *Conclusions de l'examen critique*

### *Les effets conditionnels sont-ils des "effets purs"?*

Une tendance répandue est de prendre les effets conditionnels pour des "effets propres" ou "purs" (l'idéologie de la purification !). On illustre volontiers cette terminologie par la situation de disparition. Par exemple, dans une école maternelle, on enregistre la taille des enfants et une certaine performance intellectuelle, et on trouve un effet global de la taille, mais pas d'effet de la taille conditionnellement à l'âge ; ce qui suggère la conclusion : Il n'y a pas réellement d'effet de la taille sur la performance , du moment que l'effet de la taille disparaît lorsqu'on prend en compte l'âge.

Mais dans la situation de l'*émergence*, comment peut-on avoir un effet pur alors qu'on a pas d'effet apparent ? Le cas de l'émergence pose un problème supplémentaire. Pour trouver une variable émergente dans la régression, il faut évidemment que cette variable figure parmi les variables du modèle-cadre. Si le choix des variables à retenir a été guidé par l'existence d'effets globaux, on risque de laisser de côté une variable essentielle pour l'interprétation.

Dans la hâte de trouver les effets purs («toutes choses égales par ailleurs» selon la formule rituelle), on en vient à oublier d'examiner les effets globaux qui, au moins, ne dépendent pas du modèle-cadre. Pour oser qualifier d'"effets purs" les effets conditionnels, il faut se sentir singulièrement assuré qu'on tient bien le bon modèle-cadre!

*A propos,*

*Que veut dire au juste «Toutes choses égales par ailleurs» (ceteris paribus) ?*

L'emploi de cette locution pour qualifier des effets conditionnels est un détournement de sens. Proprement employée, cette clause renvoie à des *conditions non spécifiées* (et en général non spécifiables) nécessaires à la validité d'un énoncé. Exemple : «Si le prix d'un bien diminue, la demande de ce bien augmente, *ceteris paribus*». Entendons : si le marché est "normal", sans événement perturbateur (grippe aviaire ou guerre à l'horizon, etc.). Une discipline qui admet des " lois" *ceteris paribus* peut–elle prétendre à la dignité de "science"? Problème épistémologique crucial pour l'économie (authentique science ou "première recalée" ?), traité à fond par Andler (2002, Vol 2, p.748 sq)[55].

### *Illusions et pollution*

La difficulté centrale de la régression linéaire s'étend aux diverses sortes de régression, toutes les fois qu'on a des liaisons élevées entre des variables prédictrices redondantes (ne serait-ce pas le cas, par exemple, pour le dossier Mc Fadden ?).

La *significance fallacy*, évoquée plus haut, n'est pas restreinte à la régression. Mais dans la régression, elle redouble l'illusion: un effet est tenu pour "vrai" non seulement parce qu'il est épuré ( "toutes choses égales par ailleurs") mais aussi parce qu'il est significatif. Ce *double nœud cognitif* conduit à une *sur-interprétation permanente* (et institutionnalisée) des données.

La régression est un outil utile et puissant. Mais son usage inconsidéré engendre une pollution épaisse, qui dans bien de travaux empêche de discerner les conclusions valables ![56]

[55] D. Andler, A. Fagot-Largeau, B. Saint-Sernin (2002), *Philosophie des sciences,* Gallimard, Folio-Essais.

[56] Au moins, les usagers de l'Analyse des Données qui sont réservés à l'égard de la régression auront eu le bénéfice secondaire de mettre leurs travaux à l'abri de cette pollution !

*Comment parler de "preuve statistique"?*

En statistique mathématique, l'idée de "preuve" renvoie exclusivement à la cohérence interne du discours mathématique. En statistique des chercheurs, l'administration de la preuve, évidemment vitale, apparaît avant tout affaire sociale, liée aux normes de publication. L'examen critique de la méthodologie statistique tend ainsi à devenir une préoccupation tout aussi étrangère aux " mathématiciens" qu'aux chercheurs. En somme la statistique est devenue un des ces lieux communs aristotéliciens « avec lesquels on argumente, mais sur lesquels on n'argumente pas »[57].

Plus que jamais, la question se pose d'examiner si ceux qui croient pouvoir séparer et quantifier les "effets vrais" sont ou non victimes d'une illusion. Répondre comme nous l'avons fait n'empêche pas, bien au contraire, de chercher aussi à comprendre pourquoi ils sont victimes; pourquoi Simiand ou Passeron n'ont pas été davantage écoutés. Il faudrait analyser, documents à l'appui, les usages sociaux de la statistique visant à satisfaire des «demandes subordonnées aux impératifs du profit»[58], les demandes pressantes des décideurs pharmaceutiques[59] et des bureaucraties en tous genres[60] qui ont grand besoin d'"effets vrais"…

[57] Nous reprenons la formule de Bourdieu et Wacquant (1998): Les ruses de l'impérialisme, *Actes de la Recherche en Sciences Sociales.*

[58] Cf. Bourdieu (2001), *Science de la science et réflexivité*, p.6

[59] Voir le numéro "Economie de la Recherche", *Actes de la Recherche en Sciences Sociales* (Sept. 2006)

[60] Voir l'ouvrage de M. Volle (1980) *le métier de statisticien*, qui analyse "de l'intérieur" la marge d'autonomie du champ des statistiques officielles, en ligne sur le Web : www.volle.com/ouvrages/metier