L'ANALYSE DE RÉGRESSION EN SOCIOLOGIE DE L'ÉDUCATION

Louis-André Vallet (CNRS)

Laboratoire de Sociologie Quantitative

Centre de Recherche en Économie et Statistique

Paris

« BAROUF À BOMBACH » (HENRY ROUANET, 1978)

Au cours d'un débat télévisé sur « la femme et les études scientifiques », on aborde la question de la réussite au Bac C au cours de l'année précédente.

Un premier participant fait état d'un dossier qui fournit des statistiques ville par ville. À propos de la ville de Bombach, il exhibe le tableau suivant qui fournit les nombres de succès et d'échecs au Bac C, séparément pour les garçons et pour les filles :

	Succès	*Échec*	
Garçons	***24***	***36***	***60***
Filles	***36***	***24***	***60***

De ce tableau, il déduit les proportions de réussite respectives des garçons et des filles :

Garçons : 24/60 = 0,4 ou 40 % ***Filles : 36/60 = 0,6 ou 60 %***

Il conclut donc que les filles réussissent mieux que les garçons : la différence est de 20 points de pourcentage en faveur des premières.

Mais un deuxième participant fait état d'un dossier plus détaillé, qui fournit les résultats lycée par lycée. Or, la ville de Bombach compte deux lycées – Anastase et Bénédicte – et les statistiques relatives à chacun de ces lycées sont les suivantes :

Anastase

	Succès	*Échec*	
Garçons	*15*	*35*	*50*
Filles	*1*	*9*	*10*

Bénédicte

	Succès	*Échec*	
Garçons	*9*	*1*	*10*
Filles	*35*	*15*	*50*

Bien entendu, en ajoutant case par case ces deux tableaux, on retrouve exactement le tableau présenté par le premier participant.

Or, de chacun des tableaux par lycée, le second participant déduit les proportions de réussite suivantes :

Anastase : ***Garçons :*** ***15/50 = 0,3 ou 30 %*** ***Filles :*** ***1/10 = 0,1 ou 10 %***
Bénédicte : ***Garçons :*** ***9/10 = 0,9 ou 90 %*** ***Filles :*** ***35/50 = 0,7 ou 70 %***

Le second participant conclut donc qu'à l'intérieur de chacun des deux lycées les garçons réussissent mieux que les filles. La différence des proportions de réussite est la même dans les deux lycées : 20 points de pourcentage en faveur des garçons, soit une valeur exactement opposée à celle relevée par le premier participant !

Qui a raison ?

Le premier participant a-t-il raison ?

Ou bien le second participant a-t-il raison ?

Ou bien les deux participants ont-ils simultanément raison ? Et, dans ce cas, peut-on formuler un raisonnement logique qui permette de « faire tenir ensemble » les conclusions des deux participants ?

Qu'en pensez-vous ?

Bien entendu, les deux participants ont simultanément raison.

« Barouf à Bombach » illustre donc la forme exacerbée ou paroxystique de l'effet de structure. En effet :

Dans chacun des deux lycées, les filles réussissent moins bien que les garçons. Mais le fait que les filles soient nettement sur-représentées dans le lycée où la réussite est nettement supérieure parvient, au niveau agrégé, c'est-à-dire sur l'ensemble de la ville de Bombach, à inverser la relation statistique entre sexe et réussite.

Pour bien comprendre cette notion d'effet de structure, il est utile de faire référence au beau texte de Paul F. Lazarsfeld :

« L'interprétation des relations statistiques comme procédure de recherche »,

communication présentée au congrès de l'*American Sociological Association* de Cleveland en 1946 et publiée en français dans le recueil de textes *« L'analyse empirique de la causalité »* (1966).

Le texte débute ainsi :

« LE RÔLE DES VARIABLES TESTS

Le point de départ de la présente discussion est une procédure dont l'application est quasi automatique dans la recherche empirique : quand on a constaté la présence d'une relation entre deux variables, on entreprend généralement d'analyser le rôle de variables supplémentaires. Cette procédure peut être concrétisée par un ensemble de données qui proviennent, sous une forme quelque peu stylisée, d'une série d'études sur les préférences en matière d'émissions radiophoniques.

Si on met en relation l'âge de l'informateur et le type d'émission qu'il écoute régulièrement, on observe que les personnes âgées recherchent davantage les programmes religieux et politiques. En revanche, on n'observe aucune différence entre les groupes d'âge relativement aux émissions de musique classique.

Chacun sait que l'âge est lié au niveau culturel ; en effet, en raison de l'extension croissante de l'enseignement, on observe généralement dans une communauté donnée un niveau culturel plus élevé chez les jeunes que chez les vieux. (…)

Nous raisonnons ainsi sur trois variables : âge, instruction et écoute / non écoute. Pour simplifier, nous avons réduit chaque variable à une dichotomie. Le niveau d'instruction introduit ici pour élaborer et clarifier la relation originale est appelé *variable test* (t). L'âge correspond à ce qu'on appelle conventionnellement *variable indépendante* (x) et l'écoute / non écoute en matière d'émissions radiophoniques à la *variable dépendante* (y). »

Lazarsfeld montre alors que :

- pour les émissions religieuses, la relation initiale entre âge et écoute s'atténue très fortement après contrôle du degré d'instruction (c'est l'instruction qui joue, beaucoup plus que l'âge) :

Jeunes 17 % Vieux 26 %

Degré d'instruction supérieur *Degré d'instruction inférieur*

Jeunes 9 % Vieux 11 % Jeunes 29 % Vieux 32 %

- pour les tribunes politiques, la relation initiale entre âge et écoute se trouve renforcée après contrôle du degré d'instruction (le contrôle du niveau d'instruction révèle une relation plus forte qu'il n'apparaissait initialement) :

Jeunes 34 % Vieux 45 %

Degré d'instruction supérieur *Degré d'instruction inférieur*

Jeunes 40 % Vieux 55 % Jeunes 25 % Vieux 40 %

- pour les programmes de musique classique, l'absence initiale de relation entre âge et écoute se transforme en la présence de relations de sens opposé dans les deux niveaux d'instruction (le contrôle du niveau d'instruction révèle une relation qui était cachée auparavant) :

Jeunes 30 % Vieux 29 %

Degré d'instruction supérieur *Degré d'instruction inférieur*

Jeunes 32 % Vieux 52 % Jeunes 28 % Vieux 19 %

Bref, la variable test joue un rôle très différent selon le type d'émission ! Et Lazarsfeld continue ainsi :

« SCHÉMA GÉNÉRAL D'ANALYSE

L'essence des trois exemples précédents peut être résumée dans une formule générale. (…) La relation originale (xy) est analysée en deux relations conditionnelles correspondant à chacun des niveaux d'instruction. Un symbolisme commode pour ces deux relations est (xy ; t) et (xy ; t'). (…) La structure correspondante, dont les éléments sont constitués par deux variables originales et une variable test, peut être formulée comme suit :

$$(xy) = (xy ; t) + (xy ; t') + (xt)(ty)$$

Cette formule montre que la relation originale entre x et y peut être décrite comme la somme de deux relations conditionnelles et d'un terme supplémentaire. Ce dernier est le produit de deux termes habituellement désignés comme les *relations marginales* entre le facteur test et chacune des deux variables originales. »

Et Lazarsfeld continue encore :

« Appliquons maintenant cette formule à un certain nombre de cas concrets. On a constaté que le nombre des enfants nés dans une commune donnée était en relation avec le nombre de cigognes. Il est naturellement facile d'expliquer cette curiosité : il suffit, en effet, d'introduire comme variable test, la distinction entre communes *urbaines* et communes *rurales*. On constate évidemment que la liaison entre le nombre de cigognes et celui des naissances disparaît dans chaque groupe de communes : les liaisons sont nulles. La relation primitivement observée correspond à la conjonction de deux faits : dans les communes rurales, les cigognes sont plus nombreuses et le taux de natalité plus élevé. »

Revenons maintenant à « Barouf à Bombach ». La variable x désigne le sexe ; y correspond au résultat obtenu au Bac C et le lycée constitue la variable test (avec t pour Anastase et t' pour Bénédicte) :

- dans chacun des lycées, les filles réussissent moins bien que les garçons ; les deux relations conditionnelles (xy ; t) et (xy ; t') sont donc de même sens, disons négatif (-) ;
- mais le produit des deux relations marginales est de sens opposé, disons très positif (++), car les filles sont nettement sur-représentées dans le lycée (relation (xt)) où la réussite est nettement supérieure (relation (ty)) ;
- de sorte que la somme pondérée de deux relations conditionnelles négatives et d'un produit des relations marginales très positif fournit un résultat (xy) positif (+) : dans l'ensemble de la ville de Bombach, les filles réussissent mieux que les garçons !

Plusieurs conclusions peuvent être tirées de cet exemple fictif.

1. Lorsque, dans la recherche empirique, on observe une certaine relation statistique entre deux variables, on ne peut savoir à l'avance ce qu'il va advenir de cette relation quand certaines autres variables seront prises en compte ou contrôlées. En ce sens, la recherche empirique est une grande aventure !

2. Dans la réalité sociale, il n'est pas si fréquent d'observer une situation aussi extrême que celle illustrée par « Barouf à Bombach ». Cela peut toutefois advenir, comme le montre l'exemple de l'analyse des parcours scolaires des enfants de nationalité étrangère.

3. L'analyse statistique de régression constitue la forme moderne la plus usuelle pour conduire l'analyse multivariée au sens de Lazarsfeld et interpréter les relations statistiques. Pour ce faire, il est très fécond d'utiliser des modèles emboîtés :
- dans un premier modèle, la variable d'intérêt est la *seule* variable explicative introduite ; on observe alors la relation statistique brute ou globale ;
- puis, dans des modèles ultérieurs, certaines autres variables explicatives sont progressivement ajoutées au modèle de régression ;
- et l'on peut alors observer ce que devient la relation statistique initiale lorsque certaines autres variables sont contrôlées ou tenues constantes.

L’analyse de régression peut donc être utilisée de manière descriptive, c’est-à-dire comme un outil d’analyse de moyennes ou de fréquences conditionnelles.

UN EXEMPLE

DONNÉES

Elles proviennent du panel national 1989 d’élèves du second degré (ministère de l’Éducation nationale – France).

L’échantillon étudié comprend tous les enfants :

- nés le 5 d’un mois quelconque,
- entrés en classe de sixième en septembre 1989 dans un établissement public ou privé de France métropolitaine,
- dont la famille a répondu à une enquête complémentaire au printemps 1991,
- et dont la situation a pu être suivie jusqu’à sept ans, huit ans ou neuf ans après leur entrée dans l’enseignement secondaire.

L’échantillon comprend au total 17 314 élèves.

VARIABLE DÉPENDANTE (pour mesurer la réussite dans l'enseignement secondaire)

* Obtenir le baccalauréat général ou technologique après sept ans dans l'enseignement secondaire, c'est-à-dire sans redoublement (30,9 % de l'échantillon)

VARIABLE D'INTÉRÊT

* Nationalité de l'enfant (français/étranger) (7,5 % des élèves sont étrangers)

CARACTÉRISTIQUES SOCIODÉMOGRAPHIQUES

L'information recueillie prioritairement auprès de la famille (printemps 1991) et secondairement auprès du collège (automne 1989) permet de définir onze variables qui mesurent des caractéristiques sociodémographiques des enfants et de leur famille susceptibles d'affecter la réussite scolaire :

(Ressources socio-économiques ou matérielles)

* Catégorie socioprofessionnelle du chef de famille
  (dans une nomenclature détaillée en 19 postes)
* Statut d'occupation de la mère
* Nombre moyen de personnes par pièce (utilisé comme « proxy » du revenu)

(Ressources culturelles)

* Diplôme le plus élevé de la mère (ou du seul parent en cas d'absence de la mère)
* Le fait que l'un des parents a (ou n'a pas) suivi une formation post-scolaire à son initiative
* Le fait que l'un des parents est (ou n'est pas) enseignant dans le premier degré, le second degré ou l'enseignement supérieur
* Le fait que l'enfant a (ou n'a pas) un frère ou une sœur plus âgé(e) scolarisé(e) au lycée ou dans l'enseignement supérieur

(Autres aspects objectifs de la situation familiale)

* Structure de la famille
* Nombre d'enfants dans la famille
* Rang de naissance de l'enfant
* Sexe de l'enfant

MESURE INITIALE DES PERFORMANCES SCOLAIRES

Elle utilise les scores aux épreuves standardisées de français et de mathématiques ou, s'ils ne sont pas disponibles, l'évaluation de l'élève fournie par le principal du collège en lecture, français écrit, français oral et mathématiques au moment de l'entrée en sixième (septembre 1989).

C'est une variable ordinale qui correspond à une distribution en quartiles. Une analyse préliminaire montre que, « toutes choses égales par ailleurs », les enfants étrangers ont obtenu des performances légèrement moins bonnes à l'entrée en sixième.

ASPIRATIONS SCOLAIRES DE LA FAMILLE

En utilisant les réponses à l'enquête complémentaire auprès des familles (printemps 1991), deux variables indicatrices saisissent les aspirations et attentes scolaires des parents pour leur enfant :

* Le fait que les parents souhaitent (ou ne souhaitent pas) que leur enfant poursuive ses études jusqu'à 20 ans ou plus
* Le fait que les parents disent (ou ne disent pas) qu'un diplôme de l'enseignement supérieur est le diplôme le plus utile pour trouver un emploi

Une analyse préliminaire montre que, « toutes choses égales par ailleurs », les familles des enfants étrangers expriment des aspirations scolaires plus fortes que les autres familles.

MÉTHODE

Nous partons des différences qu'un simple tri croisé révèle entre les enfants étrangers et leurs condisciples pour l'indicateur de réussite (obtention / non obtention du baccalauréat après sept ans). Puis nous analysons ces différences et évaluons le degré de réussite scolaire des enfants étrangers au moyen d'une série d'analyses de régression logistique.

* Le modèle I ne contient que la variable d'intérêt, i.e. il exprime l'effet brut de la nationalité, puis des contrôles statistiques sont introduits de façon progressive et incrémentielle :

* pour la catégorie socioprofessionnelle du chef de famille (Modèle II),

En d'autres termes, nous comparons les enfants étrangers aux autres enfants …

… dans la même catégorie socioprofessionnelle

* pour toutes les autres caractéristiques sociodémographiques (Modèle III),

… et à situation familiale et sociale identique

* pour la mesure initiale des performances scolaires (Modèle IV),

… et à niveau identique à l'entrée dans le second degré

* et pour les aspirations scolaires de la famille (Modèle V).

… et à niveau familial d'aspiration scolaire identique.

RÉSULTATS

31,8 % des entrants en sixième de nationalité française ont obtenu le baccalauréat après sept ans, c'est-à-dire sans redoublement, mais le taux correspondant est de 19,4 % parmi les élèves étrangers.

En d'autres termes, les chances de succès plutôt que d'échec sont, pour les étrangers par rapport aux Français (calcul de l'*odds ratio*) :

$$\frac{19{,}4/80{,}6}{31{,}8/68{,}2} = 0{,}5162 \qquad et \qquad Log(\,0{,}5162\,) = -0{,}66$$

OBTENIR LE BACCALAURÉAT APRÈS SEPT ANS

Modèle	I	II	III	IV	V
(Français)	-	-	-	-	-
Étranger	-0.66***	-0.10	+0.25**	+0.42***	+0.32**

(régressions logistiques estimées sur l'échantillon total)

Dans le modèle I qui inclut la seule variable de nationalité, le handicap brut des enfants étrangers est fort et hautement significatif car le coefficient est estimé à -0,66.

Dans le modèle II qui ajoute un contrôle pour la catégorie socioprofessionnelle du chef de famille, le coefficient est réduit à -0,10 et devient non significativement différent de 0. Par conséquent, les différences d'appartenance de classe entre familles étrangères et familles françaises sont responsables d'une part importante du handicap des enfants étrangers dans l'enseignement secondaire français.

De façon plus surprenante, quand l'ensemble complet de caractéristiques sociodémographiques est introduit dans le modèle de régression (modèle III), le coefficient estimé pour les élèves étrangers devient positif (+0,25) et significatif au seuil de un pour cent. La conclusion est donc que les enfants étrangers ont obtenu le baccalauréat sans redoublement *plus souvent* que les enfants français de situation familiale et sociale identique.

Dans le modèle IV qui ajoute un contrôle pour le niveau de performance scolaire à l'entrée dans l'enseignement secondaire, l'avantage « net » ou « pur » des enfants étrangers s'accroît (+0,42). Cela est cohérent avec le résultat précédent selon lequel les élèves étrangers réussissaient légèrement moins bien à l'entrée au collège que leurs condisciples français de caractéristiques sociodémographiques similaires.

Enfin, dans le modèle V qui inclut les variables mesurant les aspirations scolaires des parents, l'avantage « net » ou « pur » des enfants étrangers se réduit (+0,32). Cela suggère donc que les aspirations scolaires des familles étrangères jouent un rôle pour expliquer le degré de succès de leur enfant dans l'enseignement secondaire.

DÉCOMPOSITION DE LA RÉUSSITE DANS L'ENSEIGNEMENT SECONDAIRE

1. Réussite dans le parcours scolaire :

Admission en classe de terminale après six ans

Modèle	I	II	III	IV	V
(Français)	-	-	-	-	-
Étranger	-0.56***	-0.00	+0.35***	+0.57***	+0.45***

(régressions logistiques estimées sur l'échantillon total)

2. Réussite à l'examen en lui-même :

Obtenir le baccalauréat après sept ans

(conditionnellement à l'admission en classe de terminale après six ans)

Modèle	I	II	III	IV	V
(Français)	-	-	-	-	-
Étranger	-0.58***	-0.37*	-0.30*	-0.23	-0.23

(régressions logistiques estimées sur le sous-échantillon des élèves admis en classe de terminale après six ans)

Pour obtenir le baccalauréat après sept ans, les élèves devaient être admis en classe de terminale après six ans, c'est-à-dire sans redoublement, *et* ils devaient aussi réussir l'examen du baccalauréat à leur première tentative. Il est donc possible de scinder la probabilité du premier événement en ses deux parties élémentaires : la réussite dans le parcours scolaire *et* la réussite à l'examen *en lui-même*.

La première partie du tableau suggère que c'est dans la réussite du parcours scolaire que se forge l'avantage « net » ou « pur » des enfants étrangers. La structure d'ensemble des coefficients de régression au fil des cinq modèles ressemble étroitement à celle obtenue dans l'analyse générale, mais les coefficients positifs des modèles III à V sont légèrement plus forts et plus significatifs que leurs correspondants dans l'analyse générale. De nouveau, la réduction du coefficient positif du modèle IV au modèle V suggère qu'au sein des familles étrangères les aspirations scolaires jouent un rôle médiateur pour expliquer l'occurrence de parcours scolaires réussis chez leurs enfants.

Toutefois, pour ce qui concerne la réussite à l'examen, le résultat principal de la seconde partie du tableau est qu'il n'y apparaît aucun coefficient positif et significatif. Par conséquent, nous n'observons *jamais* que les élèves étrangers réussissent davantage l'examen du baccalauréat *en lui-même* que leurs condisciples français de caractéristiques similaires.

Ainsi, ce n'est pas dans l'examen en lui-même que l'avantage « net » ou « pur » des élèves étrangers se forme. L'avantage des élèves étrangers se constitue plutôt au long du parcours scolaire, prioritairement au collège, secondairement au lycée, et les aspirations scolaires des familles étrangères jouent un rôle dans ce processus. Des analyses complémentaires démontrent ainsi que, à la fin du collège et comparativement aux autres familles de mêmes caractéristiques sociodémographiques, les familles étrangères expriment plus souvent un vœu d'orientation de leur enfant en seconde générale et technologique.

Or, jusqu'à une date récente, l'usage de l'analyse de régression sur une telle question « n'allait pas de soi » dans la communauté sociologique française (cf. par exemple l'invocation du problème « du renne au Sahara et du chameau au pôle Nord »[1]) :

« Si la modélisation ouvre au chercheur en sciences sociales des horizons heuristiques, il reste que la clause « toutes choses égales par ailleurs » sur laquelle elle se fonde présente un risque de « sociologie fiction » redoutable (discuté notamment par Passeron, 1991). L'estimation de modèles multivariés est une fiction de raisonnement expérimental, souvent « limite », précisément parce que le raisonnement expérimental sur lequel ils reposent est évidemment très éloigné de la réalité. Si on admet sans peine qu'on peut introduire, pour expliquer les choix d'orientation, à la fois les notes scolaires et le sexe (variables corrélées), pour évaluer un effet du sexe « toutes choses égales par ailleurs » (effet net restant lui-même à expliquer), le sociologue sera plus gêné devant l'introduction simultanée de l'origine sociale et de l'origine ethnique, si on entend en déduire un effet de l'origine ethnique « toutes choses égales par ailleurs ». La quête de l'effet pur tourne ici à la sociologie fiction : dans la réalité, la distribution des niveaux d'instruction des parents (de même que la plupart de leurs caractéristiques sociales) est tout sauf égale, entre enfants français et étrangers. Le sens même de cette variable est défini dans son articulation avec d'autres » (Marie Duru-Bellat, 2002, pp. 48-49).

[1] - Il s'agit de l'argument que Maurice Halbwachs reprend de François Simiand à propos de l'application de la notion de population-type :
Halbwachs M., 1935. – « La statistique en sociologie » in *La statistique, ses applications, les problèmes qu'elle soulève,* 1944, Paris, Presses Universitaires de France, 7e semaine internationale de synthèse, pp. 113-134.

Pourtant, il n'y a pas de raison de penser que l'analyse de régression serait susceptible de produire une « sociologie fiction ».

Premièrement, les doutes exprimés à l'égard de l'analyse de régression doivent être rapprochés du fait que son entrée dans la sociologie française est récente, puisqu'elle ne s'est véritablement faite qu'avec la régression logistique, c'est-à-dire il y a moins de 20 ans.

Alors que la *path analysis*, introduite par Otis Dudley Duncan au milieu des années soixante, s'est rapidement répandue dans la sociologie américaine, il n'en est pas allé de même en France de l'analyse de dépendance, méthode pourtant très proche, développée par Raymond Boudon à la même période.

L'explication pourrait tenir au fait que les sociologues français ont toujours préféré les variables qualitatives ou catégorielles aux variables quantitatives ou numériques. Pour cette raison, le modèle de régression linéaire multiple, qui suppose que la variable dépendante est quantitative continue et forme aussi le socle de l'analyse de dépendance, leur aurait été de ce fait peu adapté.

Deuxièmement, dans le monde social, les variables sont corrélées et c'est précisément pour cette raison que l'analyse de régression est utile.

À l'inverse, elle ne serait d'aucune utilité si toutes les variables explicatives étaient orthogonales ou indépendantes car, dans ce cas, un simple tri croisé entre la variable dépendante – par exemple le résultat du parcours scolaire – et la variable d'intérêt – par exemple l'origine nationale – ferait immédiatement apparaître le lien spécifique de la seconde à la première.

Troisièmement, si une combinaison linéaire de variables explicatives permettait de reconstruire *exactement* le sous-échantillon correspondant à une modalité de la variable d'intérêt – par exemple, si une combinaison particulière de modalités des variables catégorie socioprofessionnelle du père, diplôme de la mère, taille de la famille, etc. permettait d'isoler strictement la sous-population des enfants issus de l'immigration ou celle des enfants d'origine maghrébine – alors le coefficient de régression partiel associé à *cette* modalité de la variable d'intérêt ne serait pas estimable et l'on serait dans une situation de colinéarité *stricte*.

Quatrièmement, la même proposition peut être énoncée sous une forme affaiblie.

Si une combinaison linéaire de variables explicatives permettait de reconstruire *presque exactement* le sous-échantillon correspondant à une modalité de la variable d'intérêt – par exemple, si une combinaison particulière de modalités des variables catégorie socioprofessionnelle du père, diplôme de la mère, taille de la famille, etc. permettait *presque* d'isoler la sous-population des enfants issus de l'immigration ou celle des enfants d'origine maghrébine – alors il deviendrait très difficile de « séparer les effets des variables » et cela se traduirait par des estimations très incertaines puisque les erreurs-types associées aux coefficients de régression partiels seraient très grandes.

Une conclusion importante de ces deux arguments est que, sur le plan technique, l'analyse de régression dote le sociologue de diagnostics qui lui permettent de détecter des situations limites – colinéarité stricte ou quasi-colinéarité – où l'analyse statistique ne peut délivrer de résultat robuste.

Cinquièmement, il est possible de défendre une interprétation *réaliste* des coefficients de régression partiels, c'est-à-dire argumenter l'idée suivante.

Lorsque la variable d'intérêt est assez fortement corrélée avec une autre variable explicative, alors le coefficient de régression partiel associé à une modalité particulière de la variable d'intérêt traduira surtout la situation de la population correspondante *relativement à la population de référence dotée des mêmes caractéristiques sur la seconde variable explicative*.

Ou encore, pour l'exprimer plus simplement à partir du même exemple, le coefficient de régression partiel estimé pour les élèves issus de l'immigration traduira surtout la situation de ceux-ci relativement aux élèves non issus de l'immigration et appartenant aux milieux sociaux défavorisés.

Sixièmement, raisonner à partir de données simulées, mais plausibles du point de vue de la corrélation entre les variables explicatives, permet d'illustrer la pertinence de cette interprétation réaliste des coefficients de régression partiels.

Soit un échantillon de N = 1000 élèves répartis ainsi et soumis à une épreuve de performance standardisée :

- 900 sont français, parmi lesquels
    - 600 sont d'origine sociale supérieure et obtiennent un score de 80 ;
    - 300 sont d'origine sociale populaire et obtiennent un score de 60 ;
- 100 sont étrangers, parmi lesquels
    - 10 sont d'origine sociale supérieure et obtiennent un score de 82 ;
    - 90 sont d'origine sociale populaire et obtiennent un score de 50.

Sur ces données, « l'effet » associé au fait d'être étranger (plutôt que français) vaut +2 en origine sociale supérieure, mais -10 en origine sociale populaire.

Intuitivement, on pressent que l'analyse de régression « ne se trompera pas » si elle accorde, dans ses résultats, davantage d'importance à la seconde comparaison. Dans la réalité, celle-ci nous paraît en réalité primordiale puisque, beaucoup d'élèves étrangers étant d'origine sociale populaire, c'est pour cette origine que la comparaison semble la plus solide.

De même, « l'effet » associé au fait d'être d'origine sociale supérieure (plutôt que populaire) vaut +20 parmi les Français, mais +32 parmi les étrangers.

Là encore, on pressent bien que l'analyse de régression « ne se trompera pas » si elle accorde, dans ses résultats, davantage d'importance à la première comparaison qui, pour des raisons analogues à celles vues plus haut, nous apparaît comme la plus solide.

Les résultats de la régression linéaire multiple du score de performance sur l'origine sociale et la nationalité (sans interaction entre ces deux variables) se conforment-ils à ces intuitions ?

Tel est bien le cas :

- le coefficient de régression estimé pour « étranger » (par rapport à « Français ») vaut -8.51 ; il est donc intermédiaire entre les résultats des deux comparaisons, mais fortement attiré par celle en origine populaire ;

- le coefficient de régression estimé pour « origine sociale supérieure » (par rapport à « populaire ») vaut +20.52 ; il est donc intermédiaire entre les résultats des deux comparaisons, mais fortement attiré par celle réalisée parmi les Français.

Si, sans modifier les scores ni la répartition des élèves français par origine sociale, on transforme la distribution d'origine sociale des élèves étrangers *pour la rendre encore plus extrême*, on vérifie bien par l'exemple que les coefficients de régression tendent à se rapprocher des valeurs -10 et +20 caractéristiques des comparaisons qui nous apparaissent comme les plus appropriées.

Par exemple, si les 100 élèves étrangers se répartissent en 1 élève d'origine sociale supérieure et 99 élèves d'origine sociale populaire,

les coefficients de régression estimés sont -9.84 et +20.06 .

Ou encore, si les 100 élèves étrangers sont tous d'origine sociale populaire,

les coefficients de régression estimés valent bien -10 et +20 .

Tout cela montre bien, par l'exemple, que, dans un modèle de régression où l'on incorpore à la fois l'origine sociale et l'origine nationale, le coefficient de régression partiel associé à l'origine étrangère traduit bien, d'abord et avant tout, la situation relative de ces élèves par rapport aux autres, *à l'intérieur des milieux populaires ou défavorisés*.

Ainsi, on peut soutenir une interprétation « réaliste » des coefficients de régression partiels.

CONCLUSION

Par l'utilisation d'un ensemble de modèles emboîtés, l'analyse de régression permet une *description sophistiquée* progressive où les modèles servent à résumer les caractéristiques fondamentales des données sans les déformer notablement.

L'analyse de régression est alors un outil d'interprétation des relations statistiques (au sens de Lazarsfeld). Elle invite à adopter un mode de pensée dialectique.

- « Les enfants d'immigrés comptent parmi les élèves qui encourent les plus grands risques de difficultés ou d'échec scolaires, d'orientation vers les filières peu prestigieuses du système éducatif comme de sortie précoce de celui-ci ». Voici une proposition tirée de l'estimation de modèles statistiques qui rendent compte du niveau de performance atteint ou de l'issue du parcours scolaire au moyen de l'origine nationale conçue comme variable explicative *unique*.
- « Au sein des populations défavorisées, les enfants d'immigrés sont en moyenne inscrits dans une trajectoire scolaire plus positive que les autres élèves ». Voici une proposition tirée de l'estimation de modèles statistiques qui visent à démêler l'écheveau des influences pour séparer ce qui tient en propre à l'origine nationale de ce qui relève d'autres caractéristiques objectives du milieu familial et social.

Aucune de ces deux propositions n'est plus vraie que l'autre et le fait qu'elles puissent simultanément être avancées enrichit notablement notre compréhension de la réalité sociale.