L'ABC DE L'ÉCONOMIE

Les régressions,

Rodney Ramcharan

IL EST important de savoir lire. Or les professeurs d'écoles élémentaires ont observé que la capacité d'apprendre à lire de leurs élèves tend à augmenter avec la pointure qu'ils chaussent. Dès lors, pour faciliter l'apprentissage de la lecture, les autorités devraient-elles offrir des récompenses aux savants pour qu'ils trouvent des moyens d'accroître la taille des pieds des enfants? Naturellement, le fait que la pointure et la facilité d'apprentissage augmentent en même temps ne signifie pas que des grands pieds sont la cause de la plus grande facilité d'apprentissage. Les enfants plus âgés ont les pieds plus grands, mais ils ont aussi un cerveau plus développé. C'est cette croissance naturelle des enfants qui explique l'observation simple que leur pointure et leur facilité d'apprentissage tendent à s'accroître en même temps; autrement dit, il y a une corrélation positive entre les deux. Mais il n'y a évidemment aucun rapport : ce n'est pas parce qu'on a des grands pieds qu'on lit mieux.

En économie, les corrélations sont aussi très courantes. Mais il est rarement aussi facile de savoir si la corrélation entre deux variables représente un lien de causalité. Les pays qui commercent davantage avec le reste du monde ont aussi des revenus plus élevés. Cela veut-il dire que le commerce fait augmenter les revenus? Les personnes instruites ont tendance à gagner plus d'argent. Faut-il en conclure que l'éducation est la cause de l'augmentation du revenu? Il est extrêmement important de trouver des réponses précises à ces questions. Si l'allongement de la scolarité entraînait une hausse des revenus, les autorités pourraient alors réduire la pauvreté en allouant davantage de crédits à l'éducation. Si une année supplémentaire d'études engendre une augmentation de 20.000 dollars par an du revenu, les dépenses d'éducation seraient beaucoup plus rentables que si l'augmentation n'était que de 2 dollars par an.

La magie de l'ordinateur

L'idée d'effectuer des régressions remonte au XIX^{e} siècle, mais c'est la révolution informatique du XX^{e} siècle qui a catapulté l'analyse de régression dans la stratosphère, en généralisant l'utilisation des ordinateurs personnels. Pendant les années 50 et 60, les économistes devaient faire leurs calculs à l'aide de calculatrices de bureau électromécaniques. En 1970, il fallait encore 24 heures pour obtenir le résultat d'une régression d'un laboratoire informatique central, et ce après avoir passé des heures ou des jours à perforer des fiches. Une erreur de perforation (faute d'orthographe ou valeur incorrecte), et tout était à refaire!

Pour tenter de répondre à ces questions, les économistes se servent d'un outil statistique appelé analyse de régression. Les régressions permettent de quantifier la relation entre une variable et d'autres variables dont on pense qu'elles expliquent la première; elles peuvent aussi déterminer la solidité et la fiabilité de cette relation. Aujourd'hui, les économistes n'ont aucun mal à effectuer des milliers de régressions, mais cela n'a pas toujours été le cas (voir encadré). D'ailleurs, il est difficile de trouver une étude économique empirique qui n'en contienne pas au moins une. Les régressions sont aussi largement utilisées dans d'autres domaines, dont la sociologie, la statistique et la psychologie.

De quoi s'agit-il?

Pour illustrer comment fonctionne une régression, examinons de plus près la question du rendement de l'éducation. Le gouvernement rassemble des données sur le niveau d'instruction et les revenus des personnes. Or, on peut choisir d'étudier pour des raisons très variées : certains apprennent facilement ou ont envie d'étudier plus longtemps, alors que d'autres, préférant faire carrière dans des domaines qui demandent moins d'études, gagnent quand même beaucoup d'argent. En elles-mêmes, ces motivations diverses peuvent aussi influer sur les revenus. Il est donc difficile de savoir si la corrélation entre scolarisation et revenu représente un rapport de causalité ou répond à un autre facteur. En effet, il est possible que les personnes qui apprennent facilement à l'école apprennent aussi facilement dans leur métier, d'où une augmentation de leurs revenus. Dans ce cas, la corrélation positive entre l'augmentation du revenu et le niveau d'éducation peut être due à une aptitude innée, et non aux effets de l'éducation.

Avant d'effectuer une régression, un modèle théorique peut être utile pour expliquer comment et pourquoi une variable «dépendante» est déterminée par une ou plusieurs variables «indépendantes» ou «explicatives». L'hypothèse que le revenu d'un individu dépend de son niveau d'instruction est un exemple de modèle simple à une variable explicative. L'équation correspondante, réputée linéaire, prend alors la forme suivante :

$$Y = a + bX$$

À gauche se trouve *Y*, notre variable dépendante, le revenu. À droite, nous avons *a*, une constante (le point d'interception), et *b* (la pente), multipliée par *X*, notre variable indépendante (ou explicative), l'éducation. La régression dit sous forme algébrique que «le revenu dépend uniquement de l'éducation et de façon linéaire»; les autres facteurs explicatifs éventuels ont été omis.

marotte des économistes

Mais si nous pensons que le monde est beaucoup plus complexe et que plusieurs facteurs peuvent avoir des effets sur le revenu, nous pouvons effectuer une régression à variables multiples, qui prendrait la forme suivante :

$$Y = a + b_1X_1 + b_2X_2 + \ldots$$

Nous avons maintenant, pour expliquer le revenu, plusieurs variables *X*, telles que l'aptitude, l'intelligence, l'âge, le niveau d'instruction, l'état civil, le niveau d'instruction des parents. Les coefficients *b* des variables *X* mesurent simplement les effets de chaque variable sur le revenu, les autres variables restant constantes.

Plus intelligent donc plus riche?

Essayons d'effectuer une régression en partant de la théorie que le salaire horaire (notre variable dépendante) dépend du niveau d'éducation (notre variable explicative). Supposons aussi qu'une autre variable explicative possible, l'aptitude, mesurée par les tests de *QI,* n'a aucun effet sur le salaire autre que celui qu'elle peut avoir pendant les études. Nous combinons toutes les données sur le revenu et l'éducation, et nous obtenons la régression suivante :

$$Y = 5{,}40 + 1{,}06\ EDU$$

Le coefficient *b* nous apprend qu'une année d'étude supplémentaire est associée à une augmentation de 1,06 dollar du salaire horaire. Et pour ceux qui n'ont pas fait d'études du tout ($EDU = 0$), la constante indique que le salaire moyen est de 5,40 dollars de l'heure.

Ajoutons maintenant le *QI;* autrement dit, supposons que le revenu dépend à la fois du niveau d'instruction et du *QI*. L'équation devient :

$$Y = 5{,}40 + 0{,}83EDU + 0{,}001QI$$

Nous apprenons que les individus qui ont eu les scores les plus élevés aux tests de *QI* ont aussi des salaires horaires plus élevés. De plus, si l'effet de l'éducation reste positif, il est inférieur de 27 % à ce qu'il était quand nous ne tenions pas compte du *QI* (le chiffre de 27 % est la différence entre les coefficients : 100(1,06–0,83)/0,83). Tout cela veut dire que, au départ, nous avions surestimé l'effet de l'éducation, car nous n'avions pas tenu compte du *QI,* qui est pourtant corrélé avec l'éducation.

Les écueils à éviter

Malgré tous les avantages qu'elles offrent, les régressions ne sont pas exemptes d'écueils et sont souvent utilisées à mauvais escient. Les quatre principaux écueils sont les suivants :

Les variables omises. Il faut partir d'un bon modèle théorique pour trouver les variables qui expliquent la variable dépendante. Dans le cas simple d'une régression à deux variables, il faut penser aux autres facteurs qui peuvent expliquer la variable dépendante. Dans notre exemple, même lorsque le *QI* est pris en compte, la corrélation entre niveau d'éducation et revenu peut encore être influencée par un autre facteur qui n'a pas été inclus. Autrement dit, les personnes de l'échantillon peuvent encore être différentes sous certains aspects «non observés» qui expliquent leur niveau de revenu ultérieur, peut-être par leur choix en matière d'éducation. Ainsi, des personnes issues de familles aisées peuvent avoir plus facilement accès à l'éducation, mais la richesse d'une famille peut aussi créer davantage de relations sur le marché du travail, ce qui peut se traduire par des revenus plus élevés. Il faudrait peut-être donc inclure la richesse familiale parmi les variables.

Causalité inversée. Nombre de modèles théoriques prédisent une causalité bidirectionnelle : une variable dépendante peut alors entraîner des changements pour une ou plusieurs variables explicatives. Par exemple, un revenu plus élevé peut permettre à une personne d'investir davantage dans son éducation, ce qui fera augmenter son revenu. Cela complique la manière dont il convient d'estimer les régressions et nécessite des techniques particulières.

Erreurs de mesure. Les facteurs peuvent être mal mesurés. Par exemple, l'aptitude est difficile à jauger et les tests de *QI* sont notoirement imparfaits. De ce fait, les régressions qui incluent des mesures du *QI* ne prennent peut-être pas assez en compte l'aptitude, d'où des corrélations inexactes ou biaisées entre niveau d'éducation et revenu.

Observation trop restreinte. Le coefficient de régression ne renseigne que sur la manière dont les petits, et non les grands, changements d'une variable affectent une autre variable. Par exemple, il montrera comment une faible variation du niveau d'éducation affectera le niveau de revenu, mais il ne permettra pas à l'enquêteur de tirer des conclusions générales sur les effets de grandes variations. Si tout le monde obtient son diplôme universitaire en même temps, un étudiant fraîchement diplômé ne gagnera sans doute pas beaucoup plus, car l'offre totale de jeunes diplômés se sera aussi énormément accrue. ■

Rodney Ramcharan est économiste au Département des études du FMI.