Différentes techniques statistiques de régression

François Lasnier, professeur et chercheur
Cégep de Sainte-Foy

Le présent texte vise à comparer différentes techniques de régression. Essentiellement, la régression statistique consiste à prédire une variable à partir d'une ou de plusieurs autres variables. Elle peut viser à expliquer la variabilité d'une variable dépendante par la variabilité de plusieurs autres variables indépendantes. On peut aussi se servir de la régression pour sélectionner les meilleures variables pour prédire ou expliquer une autre variable. Par exemple, on peut chercher quelles sont les meilleures variables pour expliquer la moyenne des notes au collégiale Les objectifs de cette technique sont assez clairs. Toutefois lorsqu'on doit décider quel type de régression on va utiliser pour résoudre notre problème de recherche, il n'est pas rare que les chercheurs soient embêtés par ce choix.

Problématique

Tant dans la littérature présentée dans les revues spécialisées en éducation que dans le milieu collégial, on note que les techniques de régression sont parmi les statistiques les plus utilisées. Il est donc important de s'attarder à analyser quelques problèmes relatifs à cette méthode.

Plus spécifiquement, depuis quelques années, la régression a été très utilisée pour analyser la réussite scolaire. Dans ce cas, la réussite scolaire est considérée comme variable dépendante, ou variable à prédire. On note qu'il n'y a pas de consensus parmi les chercheurs sur la façon de mesurer la réussite scolaire. Certains mesurent cette variable par la moyenne brute ou la moyenne des scores Z. D'autres la mesurent par le taux de cours réussis, la proportion d'abandons, la proportion d'échecs et autres techniques plus ou moins précises. Une constatation est évidente ; si on veut tendre vers une meilleure mesure de la réussite scolaire, on devrait chercher à la mesurer par plusieurs variables, et non seulement par une seule variable. Mais, peu importe les mesures utilisées, on veut souvent analyser une seule variable dépendante à la fois afin de mieux la comprendre et mieux l'expliquer.

Les recherches en éducation utilisent aussi plusieurs techniques statistiques pour expliquer la réussite scolaire : MANOVA, ANOVA, régression simple, régression multiple, régression par étapes, régression à entrée forcée, régression logistique, régression polynomiale et corrélation canonique. Dans cette article, on se limitera aux différentes techniques de régression.

Enfin, et c'est le coeur du problème, même en utilisant la régression pour expliquer la réussite scolaire, les résultats peuvent être très différents selon qu'on choisit tel ou tel type de régression. Quel est le type de régression qui traduit le mieux la réalité ? Comment peut-on éviter des erreurs de décision quant à notre choix parmi les différents types de régression?

Types de régression

Comme il a déjà été dit dans l'introduction, la régression a pour principal but de prédire ou d'expliquer une variable dépendante à partir d'une ou de plusieurs variables indépendantes. Examinons sommairement chaque type de régression et les conditions de leur application. Le tableau 1 présente les différents types de régression et le modèle qui schématise la relation pour chacun de ces types.

Tableau 1. Types de régression et modèles schématisant la relation entre les variables.

Régression simple	(A ⟶ X)
Régression multiple	(A+B+C ⟶ X)
Régression par étapes	(A ⟶ X , A+C ⟶ X, A+C+B ⟶ X)
Régression par entrée forcée	(A+D ⟶ X)
Régression polynomiale	($A+A^{2}+A^{3}$ ⟶ X)
Régression logistique	(A+C+B ⟶ O ou N)
Corrélation canonique	(A+C+B ⟶ X+Y+Z)

La régression simple établit la relation entre deux variables. Généralement, ce sont des variables continues (sauf pour la régression logistique, la variable dépendante sera toujours une variable continue quand on utilisera la régression).

La régression multiple applique le même modèle que la régression simple, mais elle permet en plus d'établir la relation entre plusieurs variables indépendantes et une variable dépendante.

La régression par étapes consiste, dans une première étape, à choisir parmi les variables indépendantes la meilleure variable pour prédire la variable dépendante, puis dans une seconde étape, à choisir la deuxième meilleure variable pronostique et ce, en tenant compte de la première. Le processus se répète jusqu'à ce qu'on trouve des variables ayant une relation significative avec la variable dépendante.

La régression à entrées forcées est très semblable à la régression par étapes, mais elle permet de forcer l'entrée d'une variable en premier et ce, quel que soit son niveau de signification. Cette technique est utile lorsqu'on a une

variable indépendante principale assignée par le cadre théorique.

La régression polynomiale vise à analyser si la relation entre deux variables est linéaire ou si la relation est de type exponentiel : quadratique, cubique, quartique ou autre. Par exemple, j'ai observé que pour la relation entre les notes au secondaire et au collégial, la relation est quadratique, c'est-à-dire qu'en mettant la note du secondaire au carré, ce score explique mieux la relation entre le secondaire et le collégial : si un élève est fort au secondaire, il le sera encore plus au collégial.

La régression logistique est assez différente des autres types de régression. Dans ce cas, on a une variable dépendante non continue. Elle est donc nominale ou ordinale. Elle peut être dichotomique ou trichotomique, ou même posséder quatre niveaux et plus.

Méthodologie

Pour analyser les résultats des différents types de régression, nous comparerons les résultats obtenus par chaque type de régression sur un même échantillon. Le tableau 2 présente le nombre de sujets en fonction du programme (sciences ou sciences humaines) et de la stabilité (groupes stables versus groupes réguliers). Ici la nature de l'échantillon ne nécessite pas une discussion approfondie puisque nous nous attarderons davantage aux techniques qu'à l'interprétation des résultats.

Afin de mieux comprendre et de mieux visualiser les relations entre les différentes variables, on vous présente à la figure 1, le devis de recherche qui schématise ces relations et le cadre théorique qui a servi de base à une recherche antérieure. Dans l'exemple qu'on vous présentera, notez que les variables intervenantes et la variable indépendante seront considérées comme variables pronostiques de la moyenne collégiale. La moyenne collégiale à été compilée à partir de l'ensemble des notes brutes, en excluant la note d'éducation physique. Les abandons ont été exclus de cette moyenne.

Tableau 2. Répartition des sujets en fonction de la stabilité et des programmes (n = 422).

PROGRAMMES	STABILITÉ		TOTAL
	stables	réguliers	
sciences	146	179	325
sciences humaines	50	47	97
TOTAL	196	226	422

Résultats

Le tableau 3 présente les résultats en fonction de chaque type de régression. La première colonne présente l'ensemble des variables indépendantes ou pronostiques. La variable dépendante est la moyenne au collégial.

Tableau 3. Les indices de relation entre la moyenne collégiale et les variables prédictrices, selon différentes techniques de régression.

VARIABLES	SIMPLE	MULTIPLE	PAR ÉTAPES	FORCÉE	LOGISTIQUE
	r	*Beta*	*Beta*	*Beta*	*t (3 groupes)*
stabilité	-,09			> ,02	
programme	-,01	,19	,21	,20	3,09
sexe	,16	,08		,07	
réssess	--				
résper	--		-,08	-,09	
adaptation	,24				
liens	-,08				
expression	,15				
entraide	-,02				
appartenance	-,01				
confiance	,39				
sociabilité	-,12		-,11	-,11	2,39
compétence	,45	,27	,31	,31	3,83
nb cours	,02				
MPS	,64	,60	,63	,61	8,4
MGS	,62				
R	--	,76	,74	,75	--

Examinons d'abord la régression simple. Pour la stabilité, le programme et le sexe, la relation est indiquée par un coefficient de corrélation bisériale ; pour les autres variables, c'est le coefficient de corrélation de Pearson. On note qu'avec cette technique, nous sélectionnerions la MPS, la MGS, la perception de compétence cognitive, la confiance en soi, et l'adaptation au milieu collégial comme meilleures variables pour expliquer la moyenne collégiale.
Examinons maintenant la régression multiple. La troisième colonne donne le coefficient de régression standardisé (Beta) comme indice de relation. La régression multiple considère toutes les variables indépendantes à la fois. On constate que parmi les variables déjà identifiées comme bonnes prédictrices, seules la MPS et la perception de compétence demeurent significatives. On voit apparaître deux nouvelles variables : le programme et le sexe. (Notez que la régression peut considérer des variables nominales, à condition que ces variables soient recodées en variables factices). On note que la variable MGS, malgré une corrélation de, 62, n'apparaît plus dans le modèle. Cela est dû au fait que la régression multiple tient compte de la corrélation entre les variables indépendantes. Ainsi, presque toute la variabilité expliquée par la MGS est déjà contenue dans la MPS, c'est pourquoi le modèle exclut cette variable comme variable significative.

La colonne suivante présente encore le coefficient de régression standardisé, mais cette fois obtenu par la régression par étapes. Le modèle sélectionne d'abord la meilleure variable prédictrice (MPS), puis à l'aide de la corrélation partielle, il identifie la deuxième meilleure variable. Puis, tenant compte de ces deux variables, il identifie la troisième meilleure variable prédictrice. Ici, on remarque que la sélection des variables a encore changé. Le sexe n'est plus dans le modèle, et le lieu de résidence permanente (résper) ainsi que la sociabilité sont sélectionnés. Cela est dû aux corrélations partielles, c'est-à-dire à la relation entre la

variable dépendante et les variables qui ne sont pas encore dans le modèle et ce, en considérant l'effet des variables déjà dans le modèle.

La colonne suivante nous donne encore des résultats différents, puisqu'on a utilisé la régression à entrées forcées. Étant donné que la recherche portait sur l'étude des groupes stables, on était justifié de considérer en premier la stabilité. Cette variable a donc été considérée en premier dans ce modèle. On note encore des variations par rapport aux différents coefficients de régression ; ainsi la sélection des variables significatives est modifiée.

On peut donc observer que même si le coefficient de corrélation multiple (*R*) varie très peu entre la régression multiple (,76), la régression par étapes (,74) et la régression par entrée forcée, la sélection des variables change d'une méthode à l'autre.

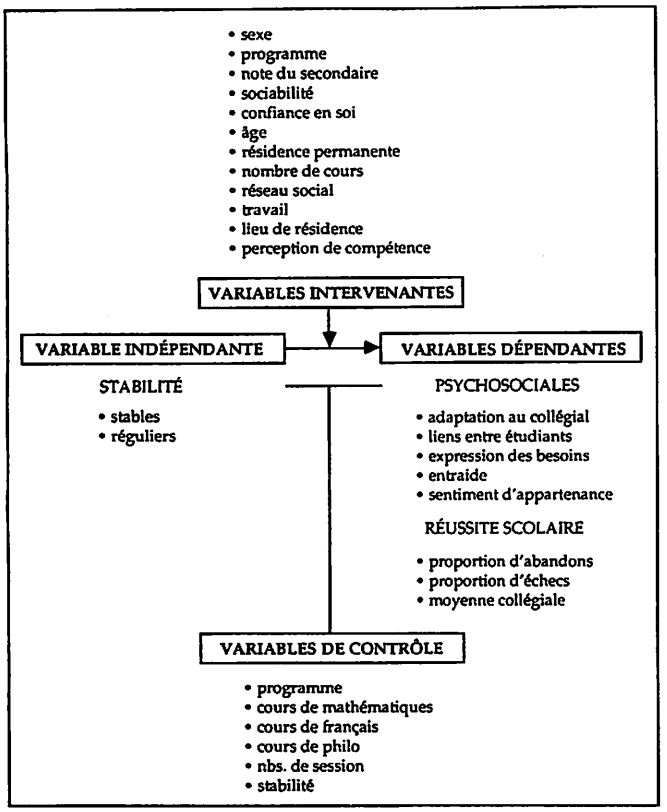

Figure 1. Devis de recherche

Finalement, on peut analyser la sélection faite à partir de la régression logistique. Ce type de régression peut considérer une variable dépendante nominale. Pour comparer cette méthode, on a donc divisé les sujets en trois groupes selon leur moyenne collégiale (faible, moyen et fort). Alors que les autres techniques de régression sont toutes basées sur la méthode des moindres carrés, la régression logistique est basée sur un modèle log-linéaire. On porte donc un jugement sur l'appartenance réelle à un groupe et son appartenance prédite. La statistique donnée dans la dernière colonne est la valeur du "t" associé à l'estimateur de chaque variable gardée dans le modèle. On note que les résultats sont semblables à ceux de d'autres régressions pour certaines variables, mais différents pour certaines.

Conclusion

L'observation la plus évidente est certainement la différence entre les méthodes. La sélection des variables pronostiques varie donc en fonction du choix de la méthode. Le problème consiste donc à faire le meilleur choix pour traduire le mieux possible une réalité.

Bien qu'il n'y ait pas de règles mathématiques formelles qui régissent ce choix, on peut donner des indices assez précis. Le choix devrait logiquement se faire en fonction des objectifs de la recherche. Généralement, le chercheur veut, soit prédire un phénomène, soit sélectionner les meilleurs indicateurs d'un concept, soit expliquer des résultats sur une variable. Chaque technique de régression a sa spécificité. Ainsi, un chercheur qui voudrait établir une méthode de sélection des élèves sur la base d'une caractéristique donnée, pourrait choisir la régression par étapes.

Cette méthode est plus économique car elle permet d'identifier un nombre réduit de variables. Cette stratégie fait donc économiser du temps de mesure. Si par ailleurs, un chercheur veut tenir compte d'un cadre théorique spécifique, il aurait avantage à utiliser la régression à entrées forcées. Mais si le chercheur veut expliquer un phénomène ou une variable spécifique, il devrait sans doute avoir recours à la régression multiple. Cette problématique est certainement la plus souvent rencontrée dans les recherches en éducation. En effet, ce qui nous intéresse le plus souvent, c'est de pouvoir expliquer un résultat en tenant compte du plus grand nombre de variables possibles à la fois. La régression multiple standard représente davantage la réalité que si on sélectionne les variables par étapes. On peut ainsi faire une analyse plus exhaustive du phénomène étudié et ce, en considérant l'effet de toutes les variables simultanément. ■