Guide pratique d'introduction à la régression en sciences sociales

Deuxième édition revue et augmentée

François Pétry
François Gélineau

Guide pratique d'introduction à la régression en sciences sociales

Deuxième édition revue et augmentée

Les Presses de l'Université Laval

Les Presses de l'Université Laval reçoivent chaque année du Conseil des Arts du Canada et de la Société d'aide au développement des entreprises culturelles du Québec une aide financière pour l'ensemble de leur programme de publication.

Nous reconnaissons l'aide financière du gouvernement du Canada par l'entremise de son Programme d'aide au développement de l'industrie de l'édition (PADIÉ) pour nos activités d'édition.

Maquette de couverture : Laurie Patry

Dépôt légal 3e trimestre 2009
ISBN 978-2-7637-8628-5

Les Presses de l'Université Laval
Pavillon Pollack, bureau 3103
2305, rue de l'Université
Université Laval, Québec
Canada, G1V 0A6

www.pulaval.com

Table des matières

Préface

Quel plaisir de voir François Pétry et François Gélineau répondre à la demande de leurs étudiants et lecteurs canadiens et européens et proposer une version revue et augmentée du *Guide pratique d'introduction à la régression en sciences sociales* publié dans la collection *Méthodes des sciences humaines* il y a quelques années. La première édition avait convaincu un grand nombre de professeurs de statistiques sociales d'adopter ce manuel pour soutenir leur enseignement de la régression dans des universités et collèges de langue française. La nouvelle édition que nous accueillons aujourd'hui confirme la place unique qu'occupe cet ouvrage dans la littérature sur les statistiques sociales.

Car il n'existe aucun autre ouvrage semblable en français. En effet, les manuels courants en français sont remplis d'un formalisme mathématique qui rebute la plupart de nos étudiants, curieux de comprendre les fondements de la statistique mais incapables de satisfaire cette curiosité sans un important investissement préalable en mathématiques.

Pétry et Gélineau relèvent avec brio le défi de rendre intelligibles les notions fondamentales de l'analyse statistique à un public dont la formation initiale en mathématiques est élémentaire. Leur recours à l'examen visuel des données cher à l'*analyse exploratoire de données* et l'économie des démonstrations mathématiques formelles rendent agréablement intuitif l'apprentissage de la régression aux étudiants et chercheurs des sciences sociales. Et cette convivialité s'accompagne de la plus grande rigueur dans la présentation des concepts et dans le traitement des données.

De plus, l'exposé de la théorie statistique s'accompagne d'applications à l'aide des deux plus importants logiciels d'analyse statistique en sciences sociales, SPSS et STATA, sur des données réelles relatives à 194 pays du monde, le tout disponible sur un site internet. Le lecteur pourra utiliser ces outils pour reproduire les résultats présentés dans chaque chapitre et faire ses propres analyses.

Cet ouvrage devrait figurer dans la bibliothèque de tous les chercheurs de langue française en sciences sociales qui sont appelés à lire des articles

scientifiques rapportant des résultats statistiques complexes, ou à utiliser de tels résultats dans la construction d'une preuve. Cet outil de référence pourrait bien leur devenir indispensable.

Louis M. Imbeau
Directeur de la collection *Méthodes des sciences humaines*

Avant-propos de la deuxième édition

Ce guide s'adresse aux étudiants et chercheurs des trois cycles qui souhaitent se familiariser avec les principales techniques de régression sans toutefois approfondir les fondements mathématiques de ces techniques au delà de ce qui est strictement nécessaire à leur bonne compréhension. Nous nous sommes efforcés de présenter le contenu en langage non technique en nous appuyant sur des exemples concrets et en privilégiant la représentation visuelle, donc plus intuitive, des relations étudiées. La lecture de ce guide ne suppose en principe aucun préalable mathématique ou statistique. Ce guide ne constitue toutefois pas, et ne vise aucunement à remplacer, un manuel statistique de base.

Cette deuxième édition revue et augmentée a été inspirée par les demandes de nos étudiants et collègues chercheurs, reflétant soit des insuffisances dans la première édition, soit des changements dans les pratiques liées à l'analyse statistique. Une première innovation concerne la présentation des commandes statistiques utilisées pour produire les tableaux et les graphiques du guide. Au lieu de présenter les menus des commandes statistiques, nous avons opté pour la présentation des syntaxes de ces commandes. Le léger supplément d'apprentissage est amplement amorti par la flexibilité supplémentaire et le gain de temps dans le traitement de données. Deuxième innovation : le lecteur n'est plus contraint d'utiliser un logiciel unique que nous imposons. Il aura désormais le choix entre les logiciels STATA et SPSS, les tableaux et les graphiques du guide étant produits par STATA. Les syntaxes ainsi que les exercices peuvent être téléchargés à l'adresse suivante : www.guide-regression.com. Cette façon de procéder allège sensiblement le guide. Dans une phase ultérieure, nous prévoyons rendre disponibles des fichiers de syntaxe pour d'autres logiciels d'analyse statistique tels que R et SAS.

La deuxième édition du guide comprend aussi des changements de contenu. Les données analysées dans le guide ont été mises à jour. Le plan du guide est toujours organisé de façon progressive, la matière y étant présentée du plus simple au plus complexe, du plus intuitif au plus formel. Le premier chapitre a été entièrement réécrit pour mieux montrer comment on conduit une recherche empirique. Le deuxième chapitre traite des concepts et des méthodes d'analyse univariée, en mettant l'accent à la

fois sur l'approche descriptive (ou exploratoire) et sur l'approche inférentielle (ou confirmatoire). Les troisième et quatrième chapitres sont consacrés à l'analyse de régression linéaire simple et multiple, respectivement. Plusieurs ajouts ont été apportés au cinquième chapitre pour permettre au lecteur de mieux comprendre les modèles autorégressifs. Le guide consacre désormais un chapitre entier (chapitre 6) à la régression logistique, permettant ainsi une explication plus détaillée de cette procédure. La deuxième édition comprend désormais un index.

Ce manuel est un guide d'introduction à la régression appliquée. Dans cette optique, nous avons choisi d'aller à l'essentiel en limitant l'exposé aux notions qui sont au centre des techniques de régression qui sont présentées sans trop nous préoccuper des concepts liés à d'autres techniques statistiques. Des notions avancées sont évoquées en fin des chapitres 4, 5 et 6 afin d'orienter les lecteurs intéressés vers des procédures statistiques plus élaborées.

Les suggestions de nos étudiants ont été très utiles pour améliorer et parfois simplifier le contenu de cette deuxième édition. Nous leurs sommes très reconnaissants. Plusieurs collègues ont lu et commenté certains ou tous les chapitres du guide et ont contribué à les améliorer. Parmi ceux-ci, nous souhaitons remercier Martial Foucault, Patrick Gonzalez, Louis Imbeau, Michael Lewis-Beck, Claude Montmarquette et Vincent Tiberj. Nous tenons particulièrement à remercier Claire Durand pour la rigueur et la pertinence de ses nombreux commentaires. L'assistance de Gabriel Blouin Genest nous a été particulièrement utile lors de l'élaboration de cette deuxième édition.

François Pétry et François Gélineau

Chapitre 1 : Comment construire une recherche empirique

1.1 INTRODUCTION

L'analyse des données n'est qu'une étape, certes importante, de la recherche empirique. Il convient donc d'étudier le processus de la recherche empirique dans son ensemble pour mieux comprendre la relation entre l'analyse des données proprement dite et les autres étapes du processus. Ce chapitre décrit sommairement les principales étapes à suivre dans la réalisation d'une recherche empirique : formulation du problème; construction du cadre opératoire; structuration de la preuve; cueillette des données; analyse des données et interprétation des résultats. Le déroulement de ces étapes sera expliqué à l'aide d'exemples qui seront d'ailleurs repris tout au long de ce guide.

Le lecteur trouvera en annexe de ce chapitre une liste de références utiles qui traitent en détail des différents aspects de la recherche empirique en général et de l'analyse des données par la technique de régression en particulier.

1.2 FORMULATION DU PROBLÈME

Toute recherche empirique en sciences sociales commence par l'identification d'un problème mettant en jeu des comportements humains. Un problème se définit comme un écart constaté entre une situation de départ insatisfaisante et une situation d'arrivée désirable. Une recherche s'entreprend afin de combler cet écart. Le chercheur est bien sûr libre d'étudier un problème de son choix selon l'intérêt qui l'anime. Il convient cependant de s'assurer dès le départ que la formulation du problème respecte deux principes généraux, à savoir que tout problème de recherche doit reposer sur un choix motivé et doit pouvoir être posé sous forme de question.

1.2.1 Un problème doit être motivé

La motivation du choix du problème à analyser peut se baser sur diverses considérations : raisons personnelles; avancement des connaissances; actualité du sujet; pertinence politique ou sociale. Pour bien motiver un problème, il convient de déterminer les utilisateurs potentiels des résultats de la recherche. Les résultats d'une recherche en sciences sociales concernent principalement deux classes d'utilisateurs potentiels : les décideurs et les autres chercheurs. Les décideurs sont surtout intéressés par l'importance sociale ou politique des résultats d'une recherche. Pour intéresser les décideurs, il convient donc de justifier une recherche par sa pertinence sociale ou politique. Les autres chercheurs s'intéressent plutôt à l'importance scientifique d'une recherche. Pour justifier une recherche aux yeux des chercheurs, il faut montrer en quoi elle contribue à l'avancement des connaissances scientifiques en corrigeant certaines lacunes trouvées dans les recherches antérieures.

Voici trois exemples de problèmes socio-politiques qui serviront aux démonstrations d'analyses tout au long de ce guide et que nous aurons donc l'occasion d'approfondir plus loin. Le premier problème est un problème de santé publique qui se pose malheureusement encore dans de nombreux pays. C'est celui de la mortalité infantile, définie comme le rapport entre le nombre d'enfants décédés à moins d'un an et l'ensemble des enfants nés vivants. Ce taux varie très fortement d'un pays à l'autre ou d'une région à l'autre, de moins de 5 pour mille dans les pays industrialisés et jusqu'à 100 pour mille et plus dans certaines régions particulièrement déshéritées du globe. Ces chiffres font clairement apparaître les coûts humains, sociaux et économiques liés au phénomène de la mortalité infantile. La mortalité infantile étant étroitement liée à la pauvreté et au manque d'hygiène des pays en développement, elle témoigne aussi des profondes inégalités qui séparent les pays en développement des pays industrialisés. L'inégalité entre pays du globe en matière de mortalité des jeunes enfants est d'autant plus tragique qu'elle pourrait être grandement et durablement soulagée à relativement peu de frais.

Un deuxième problème, économique celui-là, est celui de l'augmentation des dépenses publiques des gouvernements. La part du produit national consacré aux dépenses publiques du gouvernement canadien a plus que doublé entre 1940 et 1990, faisant craindre que la croissance de l'État soit

devenue incontrôlable. Si, comme le pensent certains économistes, le secteur public produit de façon moins efficiente que le secteur privé, la croissance incontrôlée des dépenses publiques pourrait faire craindre une baisse de la compétitivité de l'économie canadienne face aux pays dont le secteur public n'augmente pas. Plus généralement, un engrenage pervers dans lequel l'État dépense plus que l'économie nationale ne produit poserait à terme une menace sérieuse pour l'économie et les finances canadiennes.

Notre troisième problème concerne la participation électorale. C'est un problème politique très actuel puisqu'on constate que le taux de participation a constamment baissé aux élections fédérales canadiennes depuis 20 ans, et ce, surtout chez les jeunes. Le même phénomène est observé dans la plupart des démocraties industrialisées. Face à cette réalité, tant les chercheurs universitaires que les décideurs publics cherchent à mieux comprendre le phénomène. L'enjeu est de taille puisque notre système démocratique repose principalement sur les élections pour assurer aux citoyens que leurs intérêts soient représentés dans les décisions gouvernementales. Il importe donc de mieux comprendre ce problème puisque l'exclusion de certains segments de la population du processus électoral peut mettre en péril la représentation de leurs intérêts.

Jusqu'ici nous avons motivé le choix des trois problèmes qui nous préoccupent en nous basant seulement sur les aspects sociaux, politiques, économiques ou moraux de ces problèmes. Mais pour pouvoir faire l'objet d'une recherche scientifique, un problème social, politique ou moral doit à son tour être transformé en problème de recherche. Par analogie avec la définition d'un problème social, un problème de recherche se définit par l'écart constaté entre une situation de recherche insatisfaisante au départ et une situation de recherche désirable à l'arrivée. Pour établir, et ensuite combler, cet écart, il convient dans un premier temps de s'assurer que le problème choisi a fait l'objet de recherches antérieures à la fois théoriques et empiriques. Une recherche scientifique prend presque toujours racine dans des recherches antérieures, soit pour confirmer une théorie ou des résultats, soit pour les réviser ou pour les contredire. C'est sur la base des théories et des résultats produits par les chercheurs qui nous ont précédés que l'on pourra ensuite identifier les lacunes qui justifient une nouvelle

recherche. Voici une liste des principales lacunes susceptibles d'être rencontrées à l'étape de la formulation du problème de recherche :

- Mauvaise approche théorique
- Généralisations non appuyées par une démonstration empirique
- Conclusions contradictoires
- Conclusions non valides par suite de défaillances dans la méthode
- Lacunes dans la collecte des données
- Résultats de recherche périmés

1.2.2 Un problème s'énonce sous forme de question

On pose initialement une question générale de recherche que l'on précise ensuite pour arriver à une question spécifique. Par exemple, l'Organisation des nations unies (ONU) a placé la réduction de la mortalité infantile en bonne place de ses objectifs du Millénaire pour le développement. Deux facteurs souvent liés à la mortalité des enfants sont la pauvreté et le manque d'éducation. Les deux principaux leviers retenus par l'ONU pour parvenir à son objectif sont d'ailleurs l'augmentation du niveau de revenu et l'instruction, notamment chez les femmes. Mais d'autres facteurs risquent d'influencer la mortalité infantile, l'état des soins de santé en particulier. Relever le niveau de revenu et le niveau d'éducation sont bien sûr utiles, mais les résultats risquent de ne pas être optimaux s'il n'y a pas un effort parallèle d'amélioration des équipements et des services de santé publique. Dès lors il apparaît légitime de poser la question spécifique de recherche suivante : l'amélioration des services de santé publique influence-t-elle le taux de mortalité infantile ?

Examinons ensuite le problème de la croissance de l'État. L'économiste allemand Adolphe Wagner énonça voici plus d'un siècle la loi qui porte son nom selon laquelle la taille de l'État croît plus vite que le revenu national. Cela serait dû au fait que les services publics comme la sécurité publique sont des biens « supérieurs » dont la demande augmente chez les consommateurs à mesure que leurs revenus s'accroissent (ceci par

opposition aux biens de première nécessité dont l'importance relative diminue avec une augmentation du revenu national). D'autres économistes maintiennent la thèse inverse selon laquelle la croissance de l'État augmente quand l'économie nationale se contracte. Cela serait une conséquence du rôle croissant des interventions de transferts de l'État en matière de protection sociale (l'assurance emploi en particulier), transferts qui contribuent à augmenter les dépenses publiques en périodes de récession. Une fois la situation économique rétablie, le niveau des transferts sociaux baisse car il y a moins de chômeurs et la taille de l'État s'en trouve diminuée d'autant. Nous allons essayer de tirer au clair l'apparente contradiction entre ces deux interprétations en posant la question spécifique de recherche suivante : est-ce que la croissance économique a un effet positif sur l'évolution des dépenses publiques ?

Bien que les chercheurs aient étudié la participation électorale à plusieurs niveaux d'analyse, il s'agit d'un phénomène essentiellement individuel. La décision de voter ou de ne pas voter appartient ultimement à chacun des citoyens jouissant du droit de vote. Pourquoi donc les électeurs sont-ils de moins en moins enclins à se rendre aux urnes lors d'élections ? Comme nous l'avons mentionné un peu plus haut, les experts s'entendent pour affirmer que la baisse de participation est particulièrement élevée chez les jeunes. Ce faisant, certaines théories explicatives suggèrent qu'il s'agit d'un problème de socialisation des jeunes électeurs. D'autres théories affirment que la baisse du niveau de participation politique est provoquée par une hausse du cynisme à l'égard des institutions démocratiques, et ce, chez l'ensemble des électeurs. Ces différentes propositions théoriques sont-elles mutuellement exclusives ? Est-ce possible que les deux processus affectent la participation électorale ? Si oui, lequel a le plus d'impact sur le vote ?

1.3 CADRE OPÉRATOIRE

1.3.1 L'hypothèse

L'hypothèse est un énoncé déclaratif précisant une relation anticipée entre les phénomènes étudiés. En raison de son rôle central dans toute recherche empirique, il faut apporter un soin méticuleux à la formulation initiale de l'hypothèse et ne pas hésiter à la changer et à la retravailler en cours de

route le cas échéant. Lors de sa formulation initiale il faut s'assurer que l'hypothèse respecte les règles suivantes :

Une hypothèse doit être plausible, c'est-à-dire qu'elle doit avoir un rapport étroit avec le phénomène observé qu'elle prétend expliquer. Mais ce rapport n'est jamais parfait, sinon nos hypothèses démontreraient des évidences qui n'auraient plus besoin d'être vérifiées (comme par exemple l'hypothèse que l'eau gèle à 0 °C). La plausibilité d'une l'hypothèse n'est pas toujours évidente au premier regard. Pour la déterminer, il faut soigneusement observer les phénomènes étudiés parce que l'aptitude à formuler une hypothèse plausible est directement proportionnelle à la connaissance que nous aurons acquise sur l'objet d'étude. Mieux nous connaîtrons notre objet, plus nous aurons de chance de poser une hypothèse plausible à son propos. Et seule une observation approfondie des faits pertinents nous permettra de bien connaître cet objet.

Une hypothèse doit être vérifiable. Il ne sert à rien de poser une hypothèse sur le sexe des anges puisque nous ne pourrons jamais vérifier une telle hypothèse vu l'absence d'informations concrètes sur le sujet. Pour s'assurer que son hypothèse est vérifiable, un chercheur doit d'abord construire cette hypothèse sur la base de concepts (ou variables) opératoires, c'est-à-dire précis et mesurables. C'est la première condition pour pouvoir dire, à l'issue de la recherche, si l'évidence rassemblée soutient ou non l'hypothèse, la deuxième condition étant de fixer à l'avance la règle (appelée niveau de confiance) qui permettra au chercheur de dire la force avec laquelle l'évidence qu'il a cueillie soutient l'hypothèse.

Une hypothèse doit être précise. Ainsi, sa formulation doit éviter toute ambiguïté et toute confusion quant au choix des concepts ou termes clés utilisés et à la relation postulée à cette étape. Les termes clés de l'hypothèse doivent être suffisamment précis et représenter le plus adéquatement possible les phénomènes ou dimensions des phénomènes à l'étude; la relation postulée entre ces phénomènes doit aussi être spécifique et éviter toute forme d'ambiguïté.

Une hypothèse doit être générale, c'est-à-dire que son pouvoir d'explication va au-delà du cas particulier. La meilleure façon de rendre une hypothèse générale est de s'assurer qu'elle est inspirée d'une théorie causale reconnue.

Voici trois exemples d'hypothèses d'explication de la mortalité infantile, de la croissance des dépenses publiques et du déclin de la participation électorale.

- Le taux de mortalité infantile des pays du monde varie en fonction inverse du niveau de soins de santé.

- Le changement dans le niveau des dépenses du gouvernement fédéral canadien dépend de la variation dans le taux de croissance de l'économie.

- La probabilité qu'un individu vote aux élections fédérales canadiennes augmente en fonction de son âge.

1.3.2 L'unité d'analyse

L'unité d'analyse (ou unité d'observation) est l'objet ou la personne dont on étudie les caractéristiques. L'unité d'analyse, tout comme l'hypothèse, peut varier selon la stratégie de structuration de la preuve. Ainsi, dans les recherches sur la mortalité infantile et sur la croissance des dépenses publiques, l'unité d'analyse utilisée pour tester nos hypothèses est le pays alors que l'unité d'analyse dans l'hypothèse sur la participation électorale est l'individu.

À noter que les données de mortalité infantile que nous analyserons dans ce guide sont des moyennes d'observation individuelles (les enfants morts avant un an) à l'échelle de chaque pays. Le choix du pays comme unité d'analyse dans la recherche sur la mortalité infantile est un pis-aller. Choisir le pays plutôt que l'enfant comme unité d'analyse facilite la recherche en temps et en argent. Mais il y a un désavantage lié au fait que les tests statistiques d'explication de la mortalité infantile portent sur des collectivités alors que les théories causales dont ces explications s'inspirent mettent en jeu des enfants pris individuellement. Ce faisant on risque de commettre l'erreur écologique qui consiste à tirer des conclusions individuelles à partir d'évidences collectives. L'évidence statistique que les pays mieux équipés en matière sanitaire ont des taux de mortalité infantile inférieurs aux pays moins bien équipés suggère fortement que les enfants sans soins sanitaires sont plus exposés aux

risques de mortalité que les enfants bénéficiant de tels soins, surtout si cette évidence dépasse certains seuils de signification statistique. Mais cela ne prouve pas que cela soit le cas nécessairement.

1.3.3 La variable

La variable est un regroupement logique de caractéristiques décrivant un phénomène observable empiriquement. Pour éviter de confondre unité d'analyse et variable du cadre opératoire, il suffit de bien distinguer les acteurs agissants ou les objets sur lesquels ils agissent (c'est-à-dire les unités d'analyses) des attributs qui caractérisent ces acteurs (les variables). En sciences sociales, l'unité d'analyse se réduit ultimement à des individus agissants ou à des artefacts produits par des individus agissants (par exemple des pilotes d'automobiles, des travaux d'étudiants ou des acteurs de cinéma). Les variables sont les attributs qui caractérisent ces individus et objets, même si ces acteurs ou objets sont agrégés au niveau d'une collectivité, tel un pays par exemple. Dans un tableau de données statistiques, il est convenu que les unités d'analyse (les cas) sont reportées de haut en bas dans la première colonne à gauche, et que les variables sont reportées de gauche à droite dans la première rangée du haut.

Une variable peut prendre des rôles différents selon la place qu'elle occupe dans l'arrangement logique de la relation étudiée. Ici nous distinguerons les rôles distincts de variable dépendante, indépendante, variable intermédiaire et variable-contrôle.

Une variable *dépendante* est une variable dont la valeur varie en fonction de celle des autres. C'est l'effet présumé dans une relation de cause à effet et, en recherche expérimentale, c'est la variable qu'on ne manipule pas mais qu'on observe pour évaluer l'impact sur elle des changements intervenus dans les autres variables.

Une variable *indépendante* est une variable dont le changement de valeur influe sur celui de la variable dépendante. Lorsque nous postulons une relation de cause à effet, la variable indépendante est alors la cause présumée de l'effet observé. En recherche expérimentale, la variable

indépendante est la variable que le chercheur manipule pour en étudier l'influence sur la variable dépendante.

Une variable *intermédiaire* est une variable qui conditionne la relation entre la variable indépendante et la variable dépendante. C'est un élément qui permet de qualifier ou de préciser la relation reproduite dans le cadre opératoire.

Une variable intermédiaire peut parfois jouer le rôle de variable *antécédente*. Une variable antécédente est une variable qui influence à la fois la variable indépendante et la variable dépendante. Il se peut que l'influence d'une variable antécédente soit tellement dominante dans une chaîne causale qu'elle rend caduque (ou annule) la relation espérée entre la variable indépendante et la variable dépendante.

On regroupe habituellement les variables intermédiaires et les variables antécédentes sous le vocable commun de *variables-contrôle*. Comme son nom l'indique, une variable contrôle est une variable dont l'effet doit être contrôlé (en le gardant constant) dans l'examen d'une relation entre variable indépendante et variable dépendante. La recherche de variables-contrôle est un exercice incontournable lorsque l'hypothèse postule, implicitement ou explicitement, une relation de cause à effet entre deux variables.

1.3.4 L'indicateur

L'indicateur est un référent empirique de la variable dont il précise les attributs de façon à permettre la vérification empirique de l'hypothèse. La construction d'un indicateur doit respecter certaines règles. En premier lieu, l'indicateur doit obligatoirement comporter un niveau de mesure, qui varie selon la nature de la variable. On distingue trois types de niveaux de mesure : nominal, ordinal et numérique.

La mesure nominale consiste à juxtaposer les attributs de la variable sans distinction de rang, d'ordre, de proportion ou d'intervalle. Ainsi, les attributs de la variable « loi anti-déficit » sont des mesures nominales distinctes et indépendantes l'une de l'autre. Une variable nominale a soit plusieurs attributs (les métiers, ou les religions pratiquées dans un pays

donné) soit seulement deux (oui *vs* non, féminin *vs* masculin) auquel cas on l'appelle binaire ou dichotomique.

La mesure ordinale est une hiérarchisation des attributs d'une variable selon un quelconque ordre de grandeur. On pourrait par exemple classer les pays du monde en trois paquets de la variable mortalité infantile (taux élevés, taux moyens, taux faibles) pour en faire une variable ordinale.

Toutefois, puisque le taux de mortalité infantile est une variable numérique, l'utilisation d'un indicateur ordinal de cette variable perd obligatoirement de l'information et n'est donc pas recommandée sauf dans des cas particuliers. On préférera une **mesure numérique** pour la variable mortalité infantile, parce qu'elle est plus précise. La mesure numérique détermine les attributs sur la base de valeurs standardisées. Ces valeurs prennent la forme soit d'intervalles, soit de ratio. Les intervalles ont des valeurs négatives et positives. La température mesurée en degrés Celsius est une variable numérique d'intervalle. Les ratios n'ont pas de valeur négative. L'âge ou le revenu disponible sont des variables numériques de ratio.

Les indicateurs doivent également respecter les critères de précision, de fiabilité et de validité. Un indicateur doit être suffisamment précis pour permettre la réplication exacte par d'autres chercheurs. Un indicateur doit être fiable, c'est-à-dire qu'il doit donner des résultats constants dans l'espace et stables dans le temps. La validité d'un indicateur se rapporte sa capacité à représenter adéquatement le concept qu'il est censé mesurer.

Le tableau 1.1 reporte certains déterminants de la mortalité infantile et leurs indicateurs. Nous avons aussi attribué un nom court à chaque variable pour répondre aux exigences des logiciels STATA et SPSS limitant leur longueur à huit caractères au maximum.

Voyons comment notre indicateur de mortalité infantile respecte les trois critères énoncés. L'indicateur est précis. Il n'est toutefois pas très fiable puisque les données reposent sur des observations annuelles. La mortalité infantile peut être influencée indûment dans certains pays par des catastrophes ou des guerres particulièrement meurtrières, et leur effet risque d'être encore plus sévère dans les très petits pays. Pour en avoir le cœur net, le chercheur doit donc autant que possible vérifier que les taux observés dans une année donnée ne diffèrent pas sensiblement des taux

des autres années. En cas de trop grandes variations, il est préférable de prendre la moyenne sur deux ou plusieurs années consécutives.

TABLEAU 1.1 HYPOTHÈSE LIANT LA MORTALITÉ INFANTILE AUX SOINS DE SANTÉ

	Variable + nom	***Indicateur et sa mesure***
Variable dépendante	Mortalité infantile *MORTINF*	Nombre de décès avant un an pour mille naissances en 2005 (ou la plus récente année). Numérique
Variable indépendante	Soins de santé *SANTÉ*04	Dépenses en santé brutes par tête en dollars US en 2004. Numérique
1re variable contrôle	Richesse *PIBCAP*	Produit intérieur brut en dollars US par tête en 2005. Numérique
2e variable contrôle	Alphabétisation des femmes en âge de procréer *ALPHABET*	Taux d'alphabétisation des adultes en 2005 (ou la plus récente année). Numérique

L'indicateur de la variable mortalité infantile est valide dans la mesure où ce que le chercheur a besoin de mesurer dans sa recherche est bien le pourcentage d'enfants décédés avant un an. Par contre si le chercheur a plutôt besoin de connaître la proportion d'enfants qui survivent après cinq ans, l'indicateur de mortalité infantile ne sera pas valide. Il devra alors utiliser un autre indicateur plus valide : le taux de mortalité juvénile, défini comme le rapport entre le nombre d'enfants décédés à moins de cinq ans et l'ensemble des enfants nés vivants.

Le taux d'alphabétisation des adultes (*adult literacy rate*) est un bon indicateur du phénomène d'alphabétisation en général. Toutefois, notre théorie prédit que la mortalité infantile baisse en fonction d'une augmentation de l'alphabétisation des femmes en âge de procréer. L'indicateur d'alphabétisation des adultes en général inclut l'alphabétisation des hommes qui peut varier indépendamment de l'alphabétisation des femmes. Notre indicateur n'est donc pas entièrement

TABLEAU 1.2 HYPOTHÈSE LIANT LES DÉPENSES PUBLIQUES À LA CROISSANCE DU PIB

	Variable + nom	***Indicateur et sa mesure***
Variable dépendante	Dépenses publiques $DEPENSE$	Dépenses des administrations fédérale, provinciales et locales en % du produit intérieur brut. Numérique
Variable indépendante	Croissance du PIB	Augmentation du produit intérieur brut canadien d'une année sur l'autre. Numérique
Variable contrôle	Taux de chômage $CHOMAGE$	Nombre de personnes à la recherche de travail en pourcentage de la population active. Numérique

valide : il ne reflète pas toujours bien le phénomène qui nous intéresse. La banque de données de l'OMS ne contient malheureusement pas les données d'alphabétisation des femmes en âge de procréer. L'utilisation du taux d'alphabétisation des adultes en général comme indicateur est un pis-aller acceptable ici étant donné le propos didactique de ce guide. Mais il

TABLEAU 1.3 HYPOTHÈSE LIANT LE VOTE À L'ÂGE

	Variable + nom	***Indicateur et sa mesure***
Variable dépendante	Vote $VOTE$	Le fait de déclarer avoir voté. Nominale (a voté=1; n'a pas voté=0)
Variable indépendante	Âge AGE	Nombre d'années depuis la naissance. Numérique
1re variable contrôle	Revenu $REVENU$	Niveau de revenu du répondant. Ordinale (échelle en onze points de 0 à 10)
2e variable contrôle	Sexe $SEXE$	Sexe des répondants. Nominale (hommes=1; femmes=0)
3e variable contrôle	Cynisme $CYNISME$	Accord ou désaccord avec une proposition mesurant le cynisme. Ordinale (échelle en quatre points)

serait inacceptable de ne pas faire l'effort de cueillir les données d'alphabétisation des femmes en âge de procréer dans une recherche scientifique ayant pour objectif de mesurer précisément les déterminants de la mortalité infantile.

Les tableaux 1.2 et 1.3 reportent certaines variables d'explication de la croissance des dépenses gouvernementales et du vote aux élections fédérales canadiennes et leurs indicateurs.

1.4 STRUCTURATION DE LA PREUVE

En général, la recherche empirique a pour ambition d'expliquer des phénomènes humains en termes de causalité. Les chercheurs en sciences sociales évitent parfois de parler de causalité; ils utilisent alors des nuances de langage en disant qu'une variable X est « positivement associée » ou « liée » à une variable Y, ou que X « prédit » Y sans oser parler d'une relation de cause à effet sous prétexte que l'évidence empirique est trop incomplète pour établir de façon directe la causalité. C'est là une sage précaution. Il n'en demeure pas moins que c'est souvent une relation causale que les chercheurs ont à l'esprit. Dans ce cas, il est conseillé de postuler explicitement une relation causale et ensuite de présenter clairement l'évidence empirique qui nous permettra de dire si la relation observée a ou n'a pas les attributs d'une relation causale.

Il convient de souligner que l'analyse statistique en elle-même ne sert aucunement à établir la preuve de l'existence d'une relation causale entre X et Y. La preuve qu'il y a bien une relation de cause à effet entre les variables étudiées doit être soigneusement préparée (structurée) par le chercheur qui devra s'assurer que les quatre conditions ci-dessous sont respectées avant l'analyse proprement dite.

Une théorie causale existe permettant d'anticiper qu'une relation entre X et Y existe. Il convient donc de s'assurer que l'explication théorique liant la variable dépendante et la variable indépendante postule bien une relation causale.

La variable indépendante précède la variable dépendante. Autrement dit, les valeurs de X à un moment donné affectent les valeurs de Y plus

tard. Il est parfois difficile de faire précéder Y par X chronologiquement parce que les données de X et Y sont concomitantes. Dans la mesure du possible, il est recommandé d'utiliser des données de la variable indépendante qui précèdent, selon les cas, d'une semaine, d'un mois, d'une année ou même plus, les données de la variable dépendante. Il ne faut pas que les données de la variable dépendante précèdent les données de la variable indépendante. Ainsi dans l'exemple de l'explication de la variation du taux de mortalité infantile, nous utiliserons $SANTÉ04$ comme variable explicative, c'est-à-dire les dépenses de santé en 2004, soit un an avant que les données de $MORTINF$ qui sont cueillies en 2005. L'avantage est double. Non seulement on évite la confusion liée à la causalité temporelle mais en plus, l'emploi de $SANTÉ04$ plutôt que de $SANTÉ05$ se justifie du fait que les effets des dépenses de santé sur le taux de mortalité infantile prennent un certain temps à se matérialiser. Nous aurions pu choisir des données beaucoup plus espacées dans le temps (de dix années par exemple) afin de mesurer l'effet dans le long terme des dépenses de santé sur le taux de mortalité.

Les données de Y covarient avec les données de X. Cette condition n'a pas besoin d'être clarifiée plus avant sinon pour préciser que les deux variables doivent covarier de façon régulière et systématique, et non par hasard. Il existe différentes procédures statistiques pour mesurer si une covariation s'éloigne du hasard. Nous étudierons certaines de ces procédures dans les chapitres à venir.

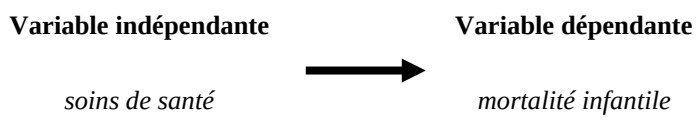
La relation entre X et Y n'est pas fallacieuse. Une relation entre X et Y est fallacieuse (ou caduque) si elle résulte de l'influence d'une troisième variable. On fait souvent appel à la recherche expérimentale pour s'assurer que la relation entre X et Y est bien réelle et non fallacieuse, comme par exemple une recherche médicale visant à tester expérimentalement l'impact d'un nouveau médicament sur les maladies cardio-vasculaires. Le chercheur divise un large échantillon de sujets en deux groupes : le groupe expérimental dans lequel les sujets reçoivent une dose quotidienne du médicament en question, et le groupe de contrôle dans lequel les sujets reçoivent un placebo. À la fin de l'expérience, le chercheur compare la fréquence des épisodes cardio-vasculaires dans chaque groupe et conclut que le médicament a un effet seulement si le nombre d'épisodes cardio-vasculaires à l'intérieur du groupe expérimental

est significativement inférieur au nombre d'épisodes à l'intérieur du groupe de contrôle.

La stratégie expérimentale se heurte malheureusement à plusieurs obstacles majeurs qui compliquent son utilisation dans les recherches en sciences humaines et sociales. En premier lieu, il faut disposer d'un large échantillon aléatoire pour contrôler, en partie au moins, les biais de sélection des sujets. Le recrutement d'un large échantillon fait monter les coûts de la recherche. Deuxièmement, en l'absence d'installations de laboratoire, il est difficile de contrôler les conditions dans lesquelles se déroule une expérience humaine. Un autre problème a trait à la difficulté d'appliquer un traitement aux sujets du groupe expérimental à leur insu. Or, si les sujets du groupe expérimental savent qu'ils sont traités de façon spéciale, ils risquent de changer le comportement que l'on cherche à tester par rapport à ce qu'ils auraient fait sans cette information. Autrement dit, il existe un problème de réactivité des sujets difficile à éliminer. Qui plus est, des problèmes incontournables d'ordre éthique sont également associés à la recherche expérimentale en sciences sociales.

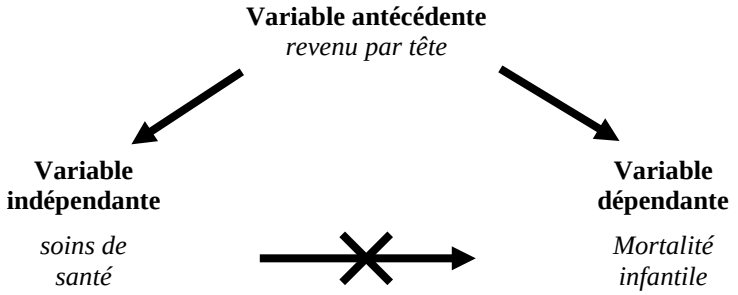
Une autre méthode beaucoup plus abordable pour s'assurer que la relation entre X et Y est bien réelle est l'enquête statistique multivariée. Cette stratégie a l'avantage d'utiliser des données d'observation documentaire dont la cueillette n'exige pas d'outillage expérimental onéreux. Elle est donc moins coûteuse et plus rapide que la méthode expérimentale. C'est d'ailleurs pourquoi c'est la stratégie la plus souvent utilisée dans la recherche empirique en sciences sociales.

Prenons un exemple pour illustrer comment une enquête statistique multivariée peut se substituer à l'expérimentation pour prouver qu'une relation causale existe entre X et Y. Supposons qu'un chercheur ait établi l'existence d'un lien négatif entre la disponibilité des soins de santé et la mortalité infantile, confirmant ainsi apparemment l'hypothèse selon laquelle la mortalité infantile diminue avec une plus grande disponibilité des soins de santé.



Supposons qu'un autre chercheur reprenne les données utilisées par le premier chercheur et introduise dans sa recherche une troisième variable, en l'occurrence le revenu par tête. En répartissant les pays étudiés selon leur niveau de revenu par tête, le deuxième chercheur trouve que le revenu par tête détermine à la fois le niveau de soins de santé et le taux de mortalité infantile. Nous dirions alors que l'ajout de la variable antécédente du revenu par tête a rendu caduque la relation causale que le premier chercheur avait trouvée entre les soins de santé et la mortalité infantile.

L'association observée par le premier chercheur entre mortalité infantile et soins de santé n'était que l'effet apparent d'une relation causale plus fondamentale liant ensemble la mortalité infantile élevée, les soins de santé limités et les bas revenus.



En réalité, nous verrons que le revenu par tête influence effectivement à la fois les soins de santé et la mortalité infantile mais pas suffisamment pour rendre caduque la relation entre les soins de santé et la mortalité infantile.

1.5 CUEILLETTE DES DONNÉES

La cueillette des données qui serviront à construire les variables d'analyse se fait par observation documentaire, en consultant par exemple les annuaires statistiques officiels ou les résultats d'enquêtes de sondages. À cette étape il faut donner les références précises des sources de données. Si elles ne sont pas trop nombreuses, les données de chaque variable peuvent aussi être reportées par l'analyste sous forme de tableau en annexe de façon à faciliter la réplication par d'autres chercheurs.

Même si nous avons placé l'étape de la cueillette des données après celles de la formulation du problème, de la construction du cadre opératoire et de la structuration de la preuve, il est évident que la bonne réalisation des étapes précédentes dépend en grande partie de l'accès aux données. Rien ne sert de définir des variables et des indicateurs si on n'a pas accès aux données pertinentes. Dans une recherche reposant sur l'observation documentaire, c'est souvent en fonction de l'accès qu'on a ou qu'on n'a pas aux données que va se construire le cadre opératoire et même parfois la question de recherche. Un principe incontournable doit guider la présentation des sources des données d'analyse. Ces sources indiquées doivent être exactes, en tout cas suffisamment précises et détaillées pour permettre à d'autres chercheurs de retrouver les mêmes données et de les analyser à leur tour s'ils le souhaitent.

Les références exactes pour trouver les données des variables évoquées dans ce chapitre sont présentées ci-dessous.

TABLEAU 1.4 DESCRIPTION ET ORIGINE DES DONNÉES DE SANTÉ

Nom de la variable dans la source	*Source internet*
MORTINF Infant mortality rate per 1000 births.	**WHOSIS (**WHO Statistical Information System**)** http://www.who.int/whosis/database/core/core_select.cfm
***SANTÉ*04** Per capita government expenditure on health at international $ rate.	
PIBCAP Gross national income per capita PPP international $.	
ALPHABET Adult literacy rate (%).	

TABLEAU 1.5 DESCRIPTION ET ORIGINE DES DONNÉES DE CROISSANCE ÉCONOMIQUE

Nom de la variable dans la source	*Source internet*
DEPENSE Dépenses des administrations fédérale, provinciales et locales en millions de $. ***CROISS*** Croissance du PIB (%). ***CHOMAGE*** Taux de chômage en % de la population active.	**Statistique Canada** (base de données : CANSIM) http://www.statcan.ca/start_f.html

Les variables d'analyse du vote aux élections fédérales sont identifiées par leur nom et par la question correspondante dans l'enquête de l'Étude électorale canadienne de 2006.

TABLEAU 1.6 DESCRIPTION ET ORIGINE DES DONNÉES DE PARTICIPATION ÉLECTORALE

Nom de la variable et question de sondage correspondante dans la source.	***Source internet***
VOTE « Avez-vous voté à cette élection ? » ***AGE*** « Quelle est votre date de naissance ? » ***REVENU*** « Quel était le revenu de votre ménage net d'impôt l'année dernière ? » ***SEXE*** Noté d'office par l'enquêteur. ***CYNISME*** Indice composé de la réponse à quatre énoncés/questions (voir note 2 du chapitre 6).	**Étude électorale canadienne 2004-2006** http://ces-eec.mcgill.ca/enquetes.html#2006

1.6 ANALYSE DES DONNÉES

La première étape d'une analyse de données empiriques consiste à faire l'étude exploratoire univariée de la variable dépendante et l'étude exploratoire bivariée de la variable dépendante et de sa principale variable d'explication. Une telle étude exploratoire vise principalement la représentation visuelle simple des données d'analyse. La place des graphiques y est particulièrement importante. On poursuit par l'exploration multivariée dans le but de développer une explication causale du phénomène examiné. Lors de la phase finale de cette étape, on confirme les résultats exploratoires par l'inférence statistique. Les notions générales d'analyse exploratoire et d'analyse confirmatoire des données sont exposées au chapitre 2.

Une fois les données des principales variables d'analyse convenablement explorées, on procède à l'étude de la relation qui les relie par voie d'analyse de régression. L'analyse de régression a pour objectif de donner une mesure chiffrée de l'effet des variables indépendantes sur la variable dépendante. Elle a aussi pour objectif de s'assurer que les effets ainsi mesurés sont statistiquement significatifs (ils ne sont pas uniquement dus au hasard) et peuvent donc être généralisés à d'autres cas ou d'autres populations. Cette dernière partie de l'analyse de régression prend la forme d'un test d'hypothèse. Sans entrer dans les détails, indiquons les principales étapes du test d'hypothèse :

Énoncer l'hypothèse nulle. Celle-ci énonce l'absence de relation entre les variables de l'hypothèse. Il est important de se souvenir que quand on cherche à prouver par voie d'enquête statistique qu'une relation existe entre X et Y, la preuve que cette relation existe consiste à rejeter, à un niveau de probabilité donné, l'hypothèse nulle qu'il n'y a pas de relation entre X et Y.

- Le taux de mortalité infantile n'est pas affecté de façon significative par les variations du niveau de soins de santé.

- Les dépenses gouvernementales ne dépendent pas de la variation dans le taux de croissance de l'économie.

- L'âge n'a aucun effet sur la probabilité individuelle de voter aux élections fédérales.

Fixer le seuil de signification statistique. Si les données indiquent que l'on peut rejeter l'hypothèse nulle, on veut être sûr de ne pas se tromper en rejetant l'hypothèse nulle alors qu'elle est vraie. Établir le seuil de signification revient à décider quel risque de se tromper on juge acceptable. Les chercheurs en sciences sociales se contentent souvent d'un seuil de signification de 5 %, ce qui veut dire qu'ils acceptent de se tromper (en rejetant l'hypothèse nulle alors qu'elle est vraie) cinq fois sur cent. D'autres chercheurs préfèrent appliquer un seuil de signification statistique de 1 %, donc plus sévère, en particulier quand le nombre de cas est élevé (plus de 1000 cas).

Procéder au test statistique pertinent et comparer ensuite le résultat calculé à l'aide du test choisi avec la valeur critique de ce test. Dans ce guide, nous allons d'abord employer la technique de régression linéaire pour laquelle le test statistique le plus couramment utilisé est le test t dont le principe sera expliqué dans un chapitre ultérieur. Nous étudierons aussi d'autres techniques de régression pour lesquelles d'autres tests s'appliquent.

Prendre la décision. Si les résultats d'analyse respectent le seuil de signification statistique que l'on s'est fixé, alors on rejette l'hypothèse nulle et on répond par l'affirmative à la question de recherche. Si le seuil de signification n'est pas respecté, on ne peut pas rejeter l'hypothèse nulle et on répond par la négative à la question de recherche.

Quatre grands types de procédures de régression seront abordés dans les chapitres qui suivent : linéaire, séries temporelles, séries spatio-temporelles, et logistique. Expliquons succinctement ces quatre procédures à l'aide d'exemples.

La régression linéaire consiste à ajuster les données d'analyse à une droite (c'est d'ailleurs pourquoi on parle d'ajustement linéaire). Dans une recherche sur les déterminants de la mortalité infantile, on peut utiliser la procédure de régression linéaire pour calculer la corrélation entre le niveau de mortalité infantile et le niveau de soins de santé dans les pays du monde à un moment donné. Appliquée à une recherche sur l'impact de la croissance économique sur les dépenses budgétaires, la procédure de régression linéaire consiste à corréler à un moment donné le niveau de dépenses des pays du monde en pourcentage de leur PIB avec leurs taux de croissance économique. Dans les deux cas, nous parlerons d'une

enquête transversale (*cross-section*) où les unités d'analyse s'observent toutes au même moment mais dans des endroits différents. La régression linéaire sur deux variables, dite régression linéaire simple, est expliquée au chapitre 3 du guide. Les notions courantes de la régression linéaire sur plusieurs variables, dite régression multivariée, sont exposées quant à elles au chapitre 4.

Une autre stratégie d'analyse de la relation entre la croissance économique et les dépenses de l'État consiste à corréler sur une longue période les changements annuels consécutifs dans les niveaux de dépenses publiques d'un même pays avec les changements annuels de son taux de croissance économique. Dans ce cas, les données auxquelles nous avons affaire sont des données de séries temporelles (*time series*). Parce qu'une série temporelle est composée de données se succédant à intervalle régulier, chaque donnée d'une série temporelle risque de subir l'influence des données qui la précèdent. Dans ce cas, on parle d'autocorrélation entre données successives. Pour des raisons que nous expliquerons plus tard dans le guide, la procédure de régression linéaire ne convient pas bien à l'analyse de données autocorrélées. Il faut alors faire appel à un autre type de régression, appelé autorégression, dont la procédure est expliquée au chapitre 5 du guide.

Une troisième stratégie consiste à cumuler l'approche transversale et l'approche chronologique en construisant un devis spatio-temporel (*time-series cross-section*) reposant sur des données de séries chronologiques dans plusieurs pays. En règle générale, la procédure linéaire et la procédure d'autocorrélation ne conviennent pas vraiment aux données spatio-temporelles. On doit donc faire appel à une procédure spatio-temporelle spéciale qui est décrite succinctement à la fin du chapitre 5 du guide.

Un autre type de procédure s'applique aux situations où la variable dépendante est une variable nominale ou catégorielle, comme dans le cas du vote à l'élection fédérale. La procédure la plus courante dans un tel cas de figure est la régression logistique qui sera exposée au chapitre 6.

1.7 LECTURES RECOMMANDÉES

La liste ci-dessous donne quelques titres d'ouvrages récents sur la méthodologie de la recherche en science sociale et la recherche empirique en particulier. Suivent quelques titres de manuels d'introduction aux statistiques destinés aux étudiants et chercheurs en sciences sociales qui ne possèdent aucune formation statistique. Les ouvrages ont été sélectionnés notamment pour leur qualité, leur accessibilité et leur pertinence pour la recherche en sciences sociales.

1.7.1 Méthodologie de la recherche

BABBIE, Earl, *The Practice of Social Research*, Belmont, Cal., Wadsworth Publishing Company, dixième édition , 2004.

GAUTHIER, Benoît, (sous la direction de), *Recherche sociale, de la problématique à la collecte des données*, Sillery, Presses de l'Université du Québec, quatrième édition, 2003.

JONES, Russel, A., *Méthodes de recherche en science humaines*, traduit et adapté par Nathalie BURNAY et Olivier SERVAIS, Bruxelles, De Boeck, 1996.

1.7.2 Manuels d'introduction aux statistiques

FOX, William, *Statistiques sociales*, traduit et adapté par Louis M. IMBEAU, Québec, Les Presses de l'Université Laval, Bruxelles, De Boek, 1999.

HOWELL, David, *Méthodes statistiques en sciences humaines*, traduit et adapté par Marylène ROGIER, Bruxelles, De Boek, 1998.

TUFTE, Edward, R., *Data Analysis for Politics and Policy*, Englewood Cliffs, N.J., Prentice-Hall, 1974.

Index