paramètres

Vincent Arel-Bundock

Analyse causale et méthodes quantitatives

Une introduction avec R, Stata et SPSS

Les Presses de l'Université de Montréal

Table des matières

PARTIE IV
Solutions

PARTIE V
Annexes

Introduction

> Nous estimons posséder la science d'une chose... quand nous croyons que nous connaissons la cause par laquelle la chose est, que nous savons que cette cause est celle de la chose, et qu'en outre il n'est pas possible que la chose soit autre qu'elle n'est. Il est évident que telle est la nature de la connaissance scientifique.
>
> *Aristote, Seconds analytiques*

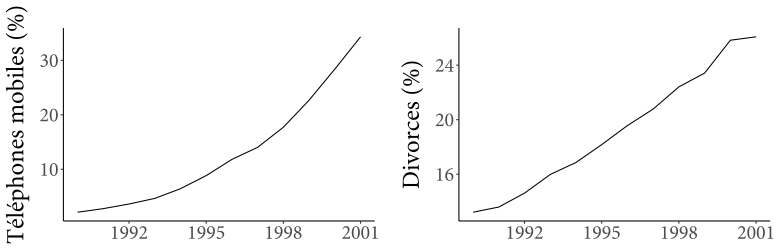
L'analyse causale est une des tâches principales du scientifique. Un criminologue évalue l'effet d'une sentence sur la probabilité qu'un condamné récidive. Une économiste mesure l'effet de la discrimination raciale sur les perspectives d'emploi d'un immigrant. Un politologue étudie l'effet des médias sociaux sur la popularité des partis d'extrême droite. Une spécialiste du marketing jauge l'effet d'une campagne publicitaire sur les choix des consommateurs.

Malheureusement, démontrer l'existence de telles relations est difficile, puisque de nombreux phénomènes sociaux ou physiques sont fortement associés, sans être liés par une relation de cause à effet. Par exemple, la figure I.1 montre que le pourcentage de la population qui utilise un téléphone mobile est fortement associé au taux de divorce : les deux phénomènes augmentent en parallèle au fil du temps et la corrélation entre eux est presque parfaite. [1] Est-ce que cette association statistique prouve que les téléphones mobiles *causent* le divorce ? Évidemment, la réponse est « non ». L'association n'implique pas la causalité.

La distinction entre association et causalité est une des pierres d'assise de la démarche scientifique. Pourtant, cette distinction est souvent ignorée dans la vie de tous les jours, quand des arguments causaux sont défendus sur la base de simples observations descriptives. Cette différence est aussi passée sous silence dans la formation méthodologique que plusieurs étudiants reçoivent à l'université. Trop souvent,

1. Le coefficient de corrélation entre les deux variables de la figure I.1 est de 0.97. Le concept de corrélation est défini au chapitre 3.

FIGURE I.1.
Taux de divorce et d'utilisation de téléphones mobiles au Canada.



Source : Banque Mondiale

Source : Statistiques Canada

les manuels de méthodes quantitatives ignorent la question causale ou recommandent d'interpréter les résultats d'un modèle statistique en termes causaux, alors qu'ils sont corrélationnels.

Pour remédier à ce problème, ce livre offre une introduction intégrée aux méthodes quantitatives et à l'analyse causale. En plus de présenter les outils nécessaires pour exécuter des analyses statistiques, il offre un cadre théorique simple et rigoureux pour interpréter les résultats de ces analyses. Ce cadre théorique permet d'identifier les conditions qui doivent être réunies afin que l'interprétation causale de nos résultats statistiques soit justifiée.

Qu'est-ce que la causalité ?

Il y a près de 300 ans, l'Écossais David Hume proposait une définition de la causalité qui guide toujours les philosophes des sciences aujourd'hui. Dans son *Enquiry Concerning Human Understanding*, Hume (1748) décrit l'analyse causale, non pas comme le fruit d'une réflexion théorique *a priori*, mais plutôt comme la conclusion qu'un analyste tire après avoir observé des régularités empiriques. Un observateur croit qu'un phénomène en cause un autre lorsque (a) la cause et l'effet sont en constante conjonction, (b) la cause et l'effet sont contigus dans le temps et l'espace et (c) la cause précède l'effet dans le temps.

Cette théorie des régularités empiriques a inspiré plusieurs philosophes, dont l'Anglais John Stuart Mill. Dans son *System of Logic*, Mill introduit plusieurs méthodes pour formaliser l'étude des régularités empiriques et pour opérationnaliser les principes de l'analyse causale.

Une des techniques les plus importantes que Mill présente dans cet ouvrage s'appelle la « méthode de différence » :

> Dans la méthode de différence il faut [...] trouver deux cas qui, semblables sous tous les autres rapports, diffèrent par la présence ou l'absence du phénomène étudié. [...] Lorsqu'un homme est frappé au cœur par une balle, c'est par cette méthode que nous connaissons que c'est le coup de fusil qui l'a tué, car il était plein de vie immédiatement avant, toutes les circonstances étant les mêmes, sauf la blessure. (Mill, 1843, Livre III, Chapitre VIII)

La méthode de différence requiert que nous puissions comparer deux cas identiques en tous points, sauf en ce qui concerne la cause qui nous intéresse. En pratique, trouver de tels cas peut être difficile. Pour Mill, et pour les générations de méthodologues qui l'ont suivi, la quête des cas comparables est un des principaux défis de l'entreprise scientifique.

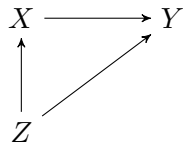
Dans les années 1970, le philosophe David Lewis a proposé une révision radicale de la théorie causale adoptée par Hume et Mill. Plutôt que de traiter l'analyse causale comme l'étude de régularités empiriques, et au lieu de chercher des individus identiques pour appliquer la méthode de différence, Lewis (1973) propose d'ancrer la causalité dans une expérience de pensée. Pour Lewis, l'analyse causale n'est pas principalement *empirique* comme chez Hume ou Mill ; il s'agit plutôt d'un exercice *théorique*. Identifier la cause d'un effet, c'est interroger un monde différent, contre-factuel et non observable. Identifier la cause d'un effet, c'est poser une question hypothétique : que se serait-il passé dans un monde contre-factuel exactement identique au nôtre, mais où la cause aurait assumé une valeur différente ?

L'approche théorique que nous adopterons dans ce livre repose sur cette expérience de pensée. Dans les chapitres qui suivent, vous serez invités à réfléchir aux mondes contre-factuels (hypothétiques) où la cause prend différentes valeurs.

Pour opérationnaliser cette réflexion contre-factuelle, et pour la lier aux techniques d'analyse statistique, nous tirerons profit de deux avancées majeures dans le champ de l'analyse causale. D'abord, des statisticiens comme Donald Rubin ont développé une nouvelle théorie statistique fondée sur l'analyse de mondes contre-factuels (chapitre 7). Cette théorie des « résultats potentiels » nous permet de mieux comprendre

les défis pratiques auxquels font face les chercheurs, et souligne l'importance des expériences aléatoires pour l'analyse causale.

En parallèle, l'ingénieur Judea Pearl développait un nouveau langage visuel qui permet d'encoder nos postulats théoriques dans de simples « graphes orientés acycliques » (chapitre 6). Par exemple, si une théorie suggère que le phénomène X cause Y, et que le phénomène Z cause X et Y, nous pourrions illustrer cette théorie ainsi :



Ce type de graphe est utile sur le plan pédagogique, parce qu'il est transparent et facile à interpréter. De plus, en appliquant quelques règles simples, Judea Pearl a démontré que les graphes orientés acycliques permettent de déterminer si les résultats d'une analyse statistique peuvent être interprétés en termes de causalité.

Grâce aux innovations de Rubin, de Pearl et de leurs collaborateurs, plusieurs disciplines vivent aujourd'hui ce que les économistes Angrist et Pischke (2010) ont qualifié de « *credibility revolution* ». L'importance de cette révolution est difficile à surestimer. En combinant la théorie de l'analyse causale et les outils des méthodes quantitatives, un chercheur peut estimer rigoureusement l'effet d'un phénomène sur un autre, et ainsi contribuer à l'accumulation des connaissances sur le monde.

Feuille de route

La première partie du livre — Analyse descriptive — introduit les lecteurs à la visualisation des données et aux principes du design graphique. Elle présente les notions de probabilités et de statistiques qui sont nécessaires pour exécuter une analyse descriptive des données et pour estimer les propriétés d'une population à partir d'un échantillon. La régression linéaire par les moindres carrés est introduite comme un outil servant à résumer l'association entre plusieurs variables. Dans cette partie du livre, tous les résultats présentés sont interprétés de façon purement descriptive, et non causale.

La deuxième partie — Analyse causale — fait le pont entre l'analyse descriptive et l'analyse causale. Elle introduit deux cadres analytiques

complémentaires : l'analyse graphique de Judea Pearl et la théorie des résultats potentiels de Donald Rubin. Ces deux approches permettent d'identifier les conditions *théoriques* qui doivent être satisfaites pour donner une interprétation causale aux résultats produits par le modèle de régression linéaire.

La troisième partie — Problèmes — est axée sur les problèmes pratiques auxquels font face les analystes qui travaillent avec des données d'observation : biais par variables omises, de sélection, de mesure et de simultanéité. Cette partie du livre souligne les principaux facteurs qui nuisent à l'inférence scientifique.

La quatrième partie — Solutions — offre aux lecteurs les outils dont ils ont besoin pour surmonter les défis de l'analyse causale. Elle explique que les expériences aléatoires sont souvent considérées comme le *Gold Standard* de l'analyse causale, parce qu'elles permettent d'éliminer plusieurs des biais identifiés dans la troisième partie du livre. Les chapitres qui suivent introduisent plusieurs techniques qui permettent de dériver des conclusions causales à partir de données d'observation : les expériences naturelles ; l'analyse de discontinuité ; l'analyse par variable instrumentale ; la méthode des doubles différences ; les modèles avec effets fixes ; les modèles multiniveaux ; le modèle linéaire généralisé ; l'analyse des effets hétérogènes ; et l'analyse de médiation. La présentation de chaque méthode est accompagnée de syntaxe informatique et de données, afin que les lecteurs puissent mettre la main à la pâte.

Finalement, l'annexe offre une introduction condensée aux concepts mathématiques et aux logiciels statistiques utilisés tout au long du livre. L'annexe présente aussi quelques concepts statistiques plus avancés.

Approche pédagogique

En écrivant ce livre, j'espère répondre aux besoins de ceux qui souhaitent acquérir une formation de base en méthodes quantitatives et en analyse causale. L'accent sur les *techniques* statistiques et sur la *théorie* causale distingue ce volume des autres textes de langue française dans le domaine.

Un autre aspect distinctif de ce livre est qu'il renvoie à beaucoup d'exemples, dont la majorité est tirée d'articles scientifiques publiés dans des revues avec évaluation par les pairs. La plupart des méthodes statistiques introduites sont mises en application en reproduisant de

vraies analyses. Par exemple, dans son exploration des expériences naturelles, le lecteur est invité à reproduire une étude sur les quotas de femmes en politique, publiée dans l'*American Political Science Review*. La méthode des doubles différences, quant à elle, est illustrée en reproduisant l'analyse d'une étude sur le salaire minimum, publiée par l'*American Economic Review*. Les lecteurs verront donc concrètement comment les méthodes quantitatives sont déployées en recherche.

Sur le plan pédagogique, ce livre innove en faisant appel à la représentation graphique des relations causales. Mon expérience suggère que les étudiants et les lecteurs répondent bien à ces graphiques. Ils sont un outil de communication efficace, qui simplifie l'exposition, complémente l'analyse algébrique et renforce l'intuition.

Ce livre diffère aussi de plusieurs manuels de statistiques, en intégrant de près les outils logiciels. Ceux-ci devront être mis à contribution par les lecteurs qui veulent compléter les exercices qui accompagnent ce volume. Mais plus encore, le logiciel statistique est intégré à la discussion au moment même où le lecteur apprend un concept ou une technique.

En vue de rendre le texte accessible au plus grand nombre, l'emploi des idées mathématiques complexes est limité au maximum. Il n'y a aucun prérequis formel pour saisir le contenu du livre. Un grand nombre d'étudiants ont excellé dans des cours développés à partir de ce livre, même s'ils n'avaient pas étudié les mathématiques depuis l'école secondaire. Le chapitre 19 en annexe présente tous les concepts mathématiques essentiels à la compréhension, ainsi que quelques idées utiles, mais non essentielles.

Pistes de lecture

À l'université, ce livre a appuyé l'enseignement de cours en méthodes quantitatives aux niveaux du baccalauréat, de la maîtrise et du doctorat. Au baccalauréat, l'enseignant pourrait encourager une lecture sélective du livre. Par exemple, les étudiants pourraient se concentrer sur les chapitres 2 à 6 et 8 à 13, en sautant les sections intitulées « Boîte à outils » ou « Analyse algébrique ». Aux cycles supérieurs, ou dans les disciplines où les étudiants ont de solides bases mathématiques, tous les chapitres devraient être accessibles à un étudiant motivé. Pour les lecteurs plus avancés, le chapitre 20 offre un traitement plus rigoureux de certains thèmes importants, dont la régression linéaire.

Logiciels statistiques : R, Stata et SPSS

Plusieurs excellents logiciels statistiques sont aujourd'hui disponibles. Ce livre est accompagné de syntaxes complètes pour trois des langages les plus populaires : `R`, `Stata` et SPSS. Pour alléger la présentation, la syntaxe du langage `R` est présentée dans le texte et les syntaxes `Stata` et SPSS se trouvent en annexe. Toutes les analyses du livre peuvent être reproduites en exécutant ces syntaxes.

Le choix d'un logiciel statistique dépend des préférences personnelles de l'analyste et ces préférences sont largement arbitraires. J'encourage donc le lecteur à utiliser le logiciel avec lequel il est le plus confiant et efficace.

Pour les lecteurs qui ne sont pas encore familiers avec un logiciel statistique, je recommande d'adopter `R`. `R` est un logiciel libre et gratuit qui a connu une hausse fulgurante de popularité au cours des dernières années. Il est en demande sur le marché de l'emploi, tant dans les secteurs privé que public. L'interface graphique `RStudio` est aussi gratuite et n'a pas d'égale dans le domaine. Finalement, les ressources pédagogiques gratuites pour `R` sont abondantes et excellentes. Une introduction au logiciel `R` est offerte au chapitre 21.

Ressources en ligne et lectures complémentaires

Le site Web qui accompagne ce livre offre plusieurs ressources. D'abord, une version électronique du livre lui-même est disponible gratuitement en version libre accès. Ensuite, toutes les banques de données utilisées dans les chapitres qui suivent sont disponibles pour téléchargement. Finalement, un ensemble étoffé de capsules vidéos, de diapositives, et d'exercices est mis à la disposition des lecteurs et des enseignants.

Un livre qui couvre autant de terrain que celui-ci doit nécessairement faire des compromis. Certains thèmes auraient mérité plus d'attention, et d'autres ont dû être laissés de côté par manque d'espace. Heureusement, d'autres auteurs ont écrit des livres complémentaires au mien.

En français, le livre édité par Gauthier et Bourgeois (2016) donne un aperçu général de la recherche en sciences sociales, de la question de recherche à la mesure, jusqu'aux méthodes qualitatives et quantitatives. Guay (2014) offre une excellente introduction au logiciel `R`, ainsi qu'à l'estimation de nombreux tests et modèles statistiques. Gélineau

(2007) et Haccoun et Cousineau (2007) offrent des traitements clairs et conventionnels des thèmes abordés dans les chapitres 2 à 5 du présent livre.

En anglais, les livres qui s'apparentent le plus à celui-ci sont ceux de Bailey (2016), Angrist et Pischke (2008), Angrist et Pischke (2014), Morgan et Winship (2014), et Cunningham (2020). Même si certains des thèmes se recoupent dans ces publications, entendre les voix de plusieurs auteurs expliquer les mêmes concepts aide à mieux comprendre.

Pour ceux qui veulent lire un traitement plus avancé et rigoureux de la théorie des résultats potentiels, je recommande Imbens et Rubin (2015). Pearl et Mackenzie (2018) et Pearl (2000) offrent des traitements détaillés de l'analyse causale par graphe orienté acyclique. [2] Hernán et Robins (2020) présentent une fusion ambitieuse des deux cadres analytiques.

Gujarati, Porter et Gunasekar (2017) et Wooldridge (2015) sont d'excellentes introductions aux méthodes quantitatives du point de vue des sciences économiques. Greene (2017) est similaire, mais plus rigoureux. Aronow et Miller (2019) adoptent une approche qui est à la fois plus fondamentale et plus moderne. Le niveau de sophistication mathématique requis pour ces deux derniers livres est plus élevé que pour le reste.

Certains manuels intègrent plus directement les logiciels statistiques à l'apprentissage. Le livre de Cameron et Trivedi (2010) couvre un large éventail de techniques statistiques et est accompagné de syntaxe `Stata`. Les livres suivants pourraient vous aider à perfectionner votre connaissance du logiciel `R` : Monogan (2015), Wickham et Grolemund (2016), Peng (2019).

Healy (2018) est un des meilleurs traitements modernes de la visualisation des données. Ce livre est accompagné d'exemples détaillés pour le logiciel `R`. Les lecteurs qui s'intéressent à la représentation de données quantitatives en cartographie pourraient se tourner vers Field (2018).

2. Le *Book of Why* de Judea Pearl vise le grand public. Il est beaucoup moins technique que *Causality*.

Pour un traitement plus approfondi des observations répétées et des données en panel, voir Wooldridge (2010). Cattaneo, Idrobo et Titiunik (2019) font une étude détaillée de l'analyse de discontinuité. Franzese et Kam (2009) traitent des effets hétérogènes. Finalement, Gandrud (2016) offre une analyse détaillée des pratiques de recherche qui favorise la robustesse et la reproductibilité des analyses quantitatives.

Remerciements

Je remercie l'extraordinaire Sari Sikilä pour ses commentaires, les exemples et pour nos longues conversations sur la démarche scientifique. Merci à Mailis et Béa Arel pour les illustrations et les questions. Merci à Evelyne, Laurent, Danièle et Charles pour l'appui sans faille et les encouragements.

Merci à André Blais, mon mentor et ami, sans qui je ne me serais pas lancé dans cette aventure. Merci à Gérard Boismenu pour sa confiance et sa vision stratégique. Florence Vallée-Dubois m'a offert des commentaires inestimables, et a eu le courage d'être la première à enseigner avec mon manuscrit. Marco Mendoza Aviña a lu ce livre plus souvent que quiconque ; il m'a offert une aide et des conseils irremplaçables.

Merci à tous les collègues, étudiants et amis qui ont contribué à ce projet. Merci à Frédérick Bastien, Laurie Beaudonnet, Charles Blattberg, Miguel Chagnon, Bill Clark, Ruth Dassonneville, David Dumouchel, Claire Durand, Rob Franzese, Jean-François Godbout, Patrick Fournier, Anne Imouza, John Jackson, Walter Mebane, Gabrielle Péloquin-Skulski, Alton BH Worthington, les étudiants des cours POL2809 et POL6021, mes collègues du département de science politique, les Bibliothèques de l'Université de Montréal, les Presses de l'Université de Montréal et deux évaluateurs anonymes.

paramètres

Ce livre offre une introduction intégrée à la théorie de l'analyse causale et aux méthodes quantitatives qui permettent d'évaluer les relations de cause à effet en sciences sociales.

Il présente les outils classiques de l'analyse descriptive (visualisation, probabilités, statistiques descriptives, inférence statistique et régression linéaire), les cadres théoriques qui facilitent le saut entre description et causalité (modèle Neyman-Rubin et graphes orientés acycliques), les défis de l'inférence causale (biais par variable omise, de sélection, de mesure et de simultanéité) ainsi que les stratégies pour les déjouer. Les exemples tirés de plusieurs disciplines en sciences sociales sont accompagnés de syntaxes informatiques complètes pour R, Stata et SPSS, et des annexes de mathématiques et de statistiques viennent ici soutenir les explications données.

Disponible en libre accès, l'ouvrage est enrichi d'un ensemble étoffé de capsules vidéo, d'exercices et de diapositives.

Vincent Arel-Bundock est professeur agrégé du Département de science politique de l'Université de Montréal.

49,95 $ • 45 €

Photo : © Konstantin Faraktinov/Shutterstock.com

Version numérique disponible en libre accès.
www.pum.umontreal.ca

ISBN 978-2-7606-4321-5