BIG DATA : PASSER D'UNE ANALYSE DE CORRÉLATION À UNE INTERPRÉTATION CAUSALE

Arthur Charpentier

Professeur d'actuariat à l'Université du Québec, Montréal

Amadou Diogo Barry

Chercheur à l'Institut de santé publique du Québec

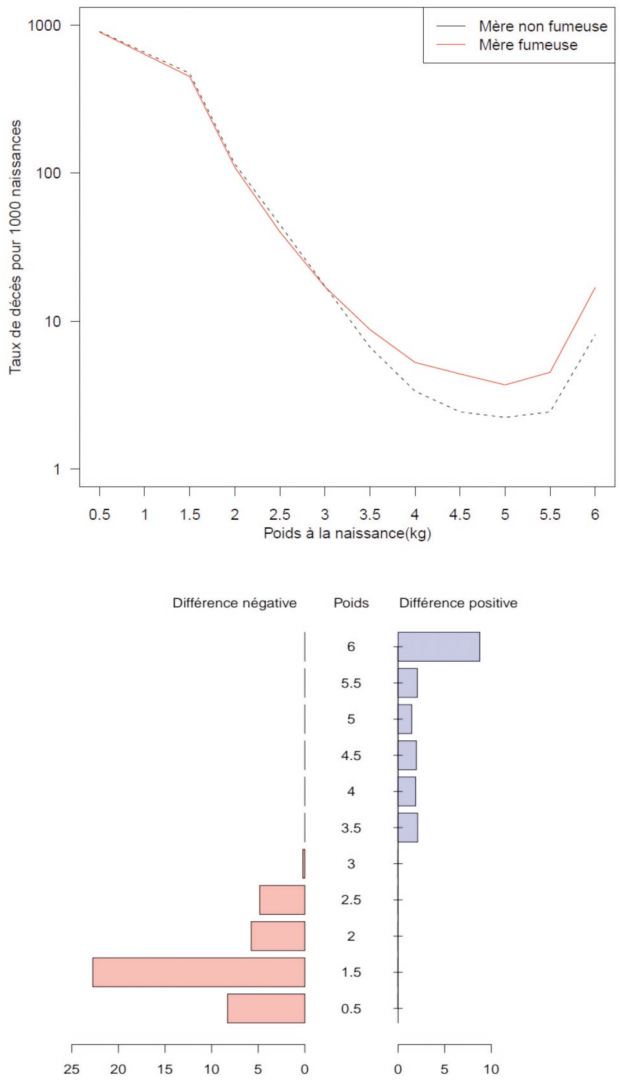
Le rôle d'un actuaire dans une entreprise d'assurance est assez souvent d'estimer la probabilité qu'un événement survienne, ou ses possibles conséquences financières, et est également fonction de variables dites « explicatives ». On voit en effet que certaines variables sont « statistiquement corrélées » avec la survenance d'un accident dans l'année, mais prétendre que l'on dispose d'une « explication » est un peu dangereux. Ce sont pourtant les interrogations qui étaient soulevées lors des débats sur la prise en compte du sexe des assurés dans la tarification automobile : si les femmes ont moins d'accident, en moyenne, que les hommes, pourquoi ne pas utiliser cette variable en tarification ? Le problème est que s'arrêter à une étude des corrélations ne permet pas de comprendre ce qui se cache vraiment derrière un phénomène. Charpentier [2014] notait que de telles études pouvaient conduire à des interprétations paradoxales et erronées.

Ainsi que le notait Dubuisson [2008], « comme le reconnaissent les actuaires, la mise en évidence d'un lien causal entre le critère choisi et la variation de la sinistralité s'apparente à la quête du Graal ». Les données massives, le big data, permettent peut-être d'avoir accès à davantage d'information et de mieux comprendre ce qui peut causer un risque. Un exemple classique est un paradoxe, long à saisir par les épidémiologistes, sur la mortalité infantile, le poids des bébés et le tabagisme de la mère. Nous allons reprendre cet exemple ici et analyser comment l'utilisation de données massives a permis de mieux comprendre ce qui pouvait réellement causer une surmortalité infantile.

Le paradoxe du tabagisme et du poids à la naissance

Le poids des bébés à la naissance est considéré comme un prédicteur important concernant la survie de l'enfant. Une autre information importante est liée au tabagisme maternel. Si l'on regarde le taux de mortalité infantile, selon le poids à la naissance, en fonction du tabagisme de la mère, on obtient le graphique de la figure 1.

Figure 1 - taux de décès (échelle logarithmique) en fonction du poids à la naissance en haut, et différence entre les taux selon que la mère est fumeuse ou pas (en bas)



Source : auteurs.

Cette figure a été obtenue à partir des données mises en libre accès sur le site du CDC (Centers for Disease Control and Prevention), contenant des informations sur toutes les naissances sur le sol américain, soit près de 4 millions d'observations par an, et plusieurs centaines de variables (dont le poids à la naissance et des informations socio-économiques sur la mère). L'analyse chiffrée porte ici sur les données de 1989.

Si on exclut les « gros bébés » (de plus de 5 kg), le taux de décès diminue à mesure que le poids augmente, ce qui est assez intuitif. Toutefois, et c'est assez surprenant, pour les bébés de poids très faible, le taux de mortalité est moindre chez ceux dont la mère fumait pendant la grossesse.

Ce graphique est une simple visualisation de probabilités conditionnelles. On représente des taux de décès « sachant que la mère fumait » et « sachant le poids de naissance ». Le danger, avec l'utilisation des probabilités conditionnelles et de la règle de Bayes, est que le « sachant » (décrivant le conditionnement) est souvent interprété de manière causale. Afin de mieux comprendre ces corrélations et ces probabilités conditionnelles, il est important de formaliser le problème. Soit la variable de tabagisme maternel, et variable indicatrice de décès. On observe que les taux de mortalité infantile aux États-Unis sont respectivement, pour une mère fumeuse,

$$\mathbb{P}(M=1|T=1)=\frac{1309}{100000}=1.31\%$$

et pour une mère non fumeuse,

$$\mathbb{P}(M=1|T=0)=\frac{864}{100000}=0.86\%$$

Un indicateur usuel pour comparer les deux risques est le « risque relatif », défini comme le ratio des deux probabilités

$$RR_{M-T}=\frac{\mathbb{P}(M=1|T=1)}{\mathbb{P}(M=1|T=0)}=1.52$$

avec un intervalle de confiance de l'ordre de [1.49 ; 1.57]. On peut aussi étudier le taux de décès en

fonction du tabagisme mais aussi du poids à la naissance. Si l'on se contente d'introduire une variable « poids trop faible » (notée P=1), on peut utiliser un modèle logistique,

$$\log \frac{\mathbb{P}(M=1|T=1)}{\mathbb{P}(M=1|T=0)} =$$

$$\beta_0 + \beta_T \mathbf{1}(T=1) + \beta_P \mathbf{1}(P=1) + \beta_{T-P} \mathbf{1}(T,P=1)$$

À partir de ces probabilités, on peut déduire deux risques relatifs : si le bébé n'est pas de poids trop faible, on retrouve un risque (significativement) plus grand que 1

$$RR_{M-T|P=0} = \frac{\mathbb{P}(M=1|T=1,P=0)}{\mathbb{P}(M=1|T=0,P=0)} = 1.72$$

avec un intervalle [1.65 ; 1.80]

ce qui correspond à notre intuition, et pour les bébé de poids très faible

$$RR_{M-T|P=1} = \frac{\mathbb{P}(M=1|T=1,P=1)}{\mathbb{P}(M=1|T=0,P=1)} = 0.78$$

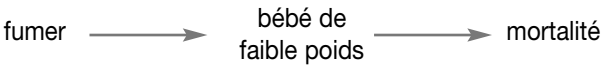
avec un intervalle [0.75 ; 0.80].

Il y a manifestement un paradoxe. Et l'avantage de disposer de tels volumes de données est de pouvoir affirmer qu'il ne s'agit pas d'un bruit statistique. Il y a statistiquement moins de risque de décès si la mère fume.

Utilisation de diagrammes causaux

La difficulté de l'exercice, que connaissent tous les actuaires, est de traduire des informations chiffrées (ici des probabilités de décéder l'année de la naissance) en une langue claire et aussi juste que possible. Les diagrammes causaux, décrits en détail dans Pearl [2000], sont aujourd'hui l'outil principal pour formaliser un tel mécanisme. Par exemple, sur la figure 2, on représente l'idée que l'on se fait de la relation causale : le fait de fumer a une influence sur le poids du bébé, lequel a lui-même une influence directe sur la mortalité. Mais il n'existe pas, dans ce schéma, de lien (direct) entre le fait que la mère fume et la mortalité du nouveau-né. Ce qui devrait se traduire, si ce modèle était juste, par le fait que si l'on considère des enfants de même poids, la probabilité de décès serait la même, que la mère ait fumé ou pas.

Figure 2 - diagramme causal 1



Source : auteurs.

À partir de cette interprétation, on peut tester ce schéma causal avec des données.

$$RR_{M-P|T=1} = \frac{\mathbb{P}(M=1|T=1,P=1)}{\mathbb{P}(M=1|T=1,P=0)} = 11.25$$

avec un intervalle [10.72 ; 11.78]

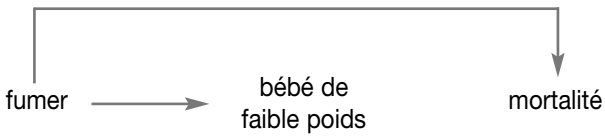
pour les mères fumeuses, alors que pour les mères non fumeuses

$$RR_{M-P|T=0} = \frac{\mathbb{P}(M=1|T=0,P=1)}{\mathbb{P}(M=1|T=0,P=0)} = 24.91$$

avec un intervalle [24.19 ; 25.63].

Ce premier diagramme causal n'est donc pas valide, compte tenu de la différence entre les risques observés : le fait que la mère fume a un impact sur la mortalité. On peut alors envisager une autre relation causale, comme sur la figure 3 : et si la mortalité infantile était en fait uniquement liée au tabagisme de la mère ? Ces diagrammes causaux se traduisent par les mêmes corrélations, mais les mécanismes en jeu ne sont pas du tout équivalents.

Figure 3 - diagramme causal 2



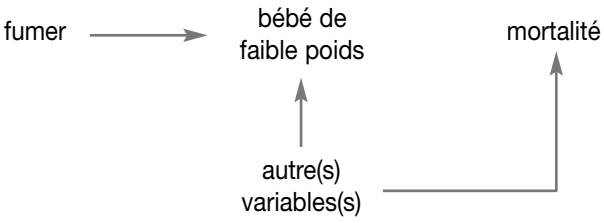
Sources : auteurs.

Les données tendent à réfuter cette idée (de causalité indirecte) affirmant que la surmortalité des bébés de faible poids serait en fait liée au fait que la mère fumait.

Recherche de causes communes et prise en compte d'autres variables

En allant un peu plus loin, on peut imaginer aussi qu'il existe des causes communes, autres que le tabagisme, entre le faible poids à la naissance et la surmortalité. Par exemple, figure 3, on peut imaginer un mécanisme causal assez simple. Fumer a un impact sur le poids des bébés (ce point est relativement bien établi par une multitude d'études), mais d'autres variables interviennent aussi (comme la sous-alimentation de la mère ou une malformation congénitale du bébé – pour simplifier, nous retiendrons cette dernière hypothèse).

Figure 4 - diagramme causal 3



Source : auteurs.

Dans ce modèle, un enfant de faible poids à la naissance dont la mère ne fume pas a forcément une malformation congénitale (c'est le principe de ces relations causales) car on a ici seulement deux causes possibles. De plus,

- une malformation congénitale augmente la mortalité infantile ;
- le tabagisme n'augmente pas la mortalité (il affecte juste le poids des bébés).

Avec ce modèle, les bébés de poids faible dont la mère ne fume pas ont alors forcément une maladie congénitale, et donc leur taux de survie diminue. On a alors une relation (corrélation) négative entre le tabagisme de la mère et la mortalité infantile. Et il n'est alors pas impossible d'avoir $RR_{(M-T|P=1)}$ inférieur à 1, comme nous l'avions observé numériquement. Les diagrammes causaux permettent de comprendre ces paradoxes. Mais reste à les tester...

Big data et modèles de médiation

On dispose de plus en plus de gros volumes de données, en particulier sur les naissances. Avec des bases non seulement exhaustives (comprenant ici toutes les naissances recensées) mais de plus en plus détaillées, et énormément d'informations sur la famille, sur l'accouchement, etc. En épidémiologie, des modèles dits de médiation – tenant compte d'informations additionnelles – ont connu un développement important ces dernières années. On peut alors calculer des taux de risques, lorsque la variable de médiation U est binaire. Le principe, dans notre dernier modèle causal, est que seule cette variable U influence la mortalité.

Autrement dit, le facteur de risque

$$RR_{M-U|T,P} = \frac{\mathbb{P}(M=1|T=t,P=p,U=1)}{\mathbb{P}(M=1|T=t,P=p,U=0)}$$

ne doit dépendre ni de t, ni de p. On peut alors tâtonner parmi toutes les variables pour en trouver qui vérifient cette propriété. Il n'est alors pas rare de voir un facteur de risque corrigé, afin de synthétiser l'information. Par exemple, $RR_{M-U|T,P}$ deviendra

$$RR^{\star}_{M-T|P} = \frac{RR_{M-T|P}}{\text{biais}(P)} \quad \text{avec}$$

$$\text{biais}(P) = \frac{1+[\gamma-1]\mathbb{P}(U=1|T=1,P)}{1+[\gamma-1]\mathbb{P}(U=1|T=0,P)}$$

Aller (bien) au-delà de la corrélation

Ces modèles permettent enfin de mieux comprendre les mécanismes causaux, et de mieux cibler les actions de prévention des instituts de santé publique. Et cela est possible grâce aux volumes colossaux de données qui sont aujourd'hui collectés, avec non seulement des bases quasiment exhaustives, contenant énormément d'observations (on retrouve le fameux n = tout le monde de Meyer-Schönberger & Cukier [2013], mais également de plus en plus de variables, souvent bien renseignées. Et si le danger des *spurious regression* rôde (régression fallacieuse, venant du fait que, parmi de nombreuses variables, on peut toujours en trouver deux parfaitement corrélées), la construction et la validation de diagrammes causaux permettent justement d'avoir des interprétations justes.

Denuit [2005] notait que « l'assureur qui désire faire usage d'un critère de segmentation doit pouvoir démontrer, statistiques à l'appui, le lien causal entre ce critère et les variations de la sinistralité qu'il est supposé induire ». Pour l'instant, certaines variables sont utilisées à cause de la corrélation apparente qui existe, et parce que les vrais facteurs de risque (dans le cas du risque automobile, agressivité au volant, non-respect du code de la route, consommation d'alcool, fatigue, etc.) ne sont pas observables et ne peuvent être incorporés dans le tarif. Avoir accès à des données plus fines (et nettement plus volumineuses) permettrait déjà de mieux comprendre les relations causales, et d'envisager d'autres critères de tarification, évitant ainsi l'iniquité inhérente à l'utilisation de variables simplement corrélées avec la sinistralité.

Bibliographie

CDC, bases de données ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Datasets/DVS/periodlinkedus/

Charpentier A., « Interprétation, intuition et probabilités », *Risques*, n° 99, septembre 2014.

Denuit M., « Quand la différenciation tarifaire est-elle techniquement justifiée ? », *Le Monde de l'assurance*, dossier spécial, 16-31 mai 2005.

Dubuisson B., « Solidarité, segmentation et discrimination en assurances, nouveau débat, nouvelles questions », 2008, http://goo.gl/LZRFXT.

Hernandez-Diaz S. ; Schisterman E. F. ; Hern MA., "The Birth Weight 'Paradox' Uncovered?", *American Journal of Epidemiology*, n° 164, 2006, pp. 1115–1120.

Meyer-Schönberger V. ; Cukier K., *Big Data: A Revolution That Will Transform How We Live, Work, and Think*, Eamon Dolan, 2013.

Pearl J., *Causality: Models, Reasoning and Inference*, Cambridge University Press, 2000.

VanderWeele T. J., "Bias Formulas for Sensitivity Analysis for Direct and Indirect Effects", *Epidemiology*, n° 21(4), 2010, pp. 540-551.