Can J Anesth/J Can Anesth (2017) 64:461–470
DOI 10.1007/s12630-017-0833-0

EDITORIALS

How to construct regression models for observational studies (and how NOT to do it!)

Kevin E. Thorpe, MMath

Received: 6 November 2016 / Revised: 14 December 2016 / Accepted: 25 January 2017 / Published online: 24 February 2017

Introduction

All models are wrong, but some are useful

This pithy aphorism, in various forms, is attributed to the eminent statistician, George Box. Despite its apparent simplicity, it conveys a number of deep issues. First, it reminds us that all statistical analyses are in some sense attempts to model real world phenomena. Second, it asserts that no statistical model is perfect and therefore must be wrong in some sense. Models can be imperfect for various reasons, such as incorrect assumptions, improper specification, and flawed model building strategies (among others). Finally, we are reminded that, despite the imperfect models we can build, some can be useful. Although the usefulness of a model is closely related to the primary goal of an analysis, for most of the discussion that follows, a *useful* model is defined as one that informs us about the population with minimal distortion and permits the researcher to generalize from a specific sample to the population with reasonable comfort. An important corollary is that some models are not useful. It should come as no surprise that good methodology is an essential ingredient when producing useful models. Unfortunately, much model building seen in the literature continues to employ flawed methodology. The remainder of this article discusses some of these flaws and presents alternatives. To provide context, we begin with a hypothetical scenario.

Two anesthesia researchers are interested in studying the relation between smoking and both perioperative complications and hospital length of stay. They both have access to the outcomes of interest via electronic records, patient smoking status, and other potentially important patient characteristics such as age, sex, type of anesthesia, etc. for thousands of patients. They plan to use logistic regression to model the presence of a complication and linear regression to model the hospital length of stay.[A] Researcher A plans to begin with univariate pre-screening. That is, each variable of interest will be tested for an association with the outcome. Any variables where $P < 0.05$ then carry forward to the next phase. During the second phase, all variables that passed the initial screening will be entered into a model, and backward elimination will be used to arrive at a final set of *statistically significant* variables. Taking a completely different approach, researcher B begins by identifying variables known, or strongly suspected, to be related to smoking status and outcome based on the available literature or theoretical grounds. Next, all identified variables are included in the model and retained irrespective of their *P*-values.

Given these two contrasting approaches, it is likely that, despite the popularity of approach A in the literature, many researchers would have a nagging suspicion that approach B is more likely to result in a useful model (in the sense defined above). In truth, approach A (and others like it) should be avoided, and approach B should be the method of

K. E. Thorpe, MMath (✉)
Dalla Lana School of Public Health, University of Toronto, Toronto, ON, Canada
e-mail: kevin.thorpe@utoronto.ca

K. E. Thorpe, MMath
Applied Health Research Centre (AHRC), Li Ka Shing Knowledge Institute of St. Michael's Hospital, Toronto, ON, Canada

[A] Linear regression is often not appropriate for assessing hospital length of stay, but we assume that it works here.

choice for data analysis. The remainder of this editorial will examine the flaws in approach A and provide advice about using approach B to obtain useful models.

Goals of models

Before proceeding with an analysis of the two approaches presented, it is worth reviewing briefly two goals of statistical models: prediction *vs* explanation. Although there are some commonalities between these two approaches, there are issues that should be approached with concern.

Explanatory models

The goal of explanatory models is understanding. That is, one wishes to understand and quantify the effects of variables on an outcome. Thus, you want effect estimates to be (ideally) free of bias. You need to account for variables that confound (i.e., mask) a relationship and guard against multicollinearity, but variables that purely *predict* the outcome are less important to take into account (see Table for explanations of these and other statistical terms). The explanatory model is likely the most common model fit for observational data. As the goal in the hypothetical scenario is to understand how smoking is related to certain outcomes, it fits the explanatory paradigm.

Predictive models

The goal of predictive models is to predict an outcome *accurately*. That is, the model should be able to predict responses well in the population. In this type of model, predictors are important, but arriving at explanations of variable associations is not important. Even multicollinearity may not be a big issue, provided the model has good predictive properties. It is important to realize that the model that fits the data *best* is likely not the best predictive model, as overfitting (Table) is likely to exist in the *best* model. If the researchers in the hypothetical scenario were interested in being able to predict reliably which patients will develop postoperative complications based on their pre-surgery characteristics, they would need to develop a predictive model. Predictive models are not considered further here. It is worth noting that a useful predictive model is one that "accurately" predicts outcomes for future subjects. It is less concerned with obtaining useful explanations of underlying relations so long as prediction works well.

Exploratory models

It may be the case that relatively little is known about a particular problem. In those situations, models may be used to form hypotheses about potential relations that need to be tested in subsequent studies. If researchers A and B were simply interested in learning what factors might be related to the outcomes of interest, exploratory models would be indicated. In fact, embodied in the approach of researcher A is exploratory thinking. The goal of exploratory models may be eventually to arrive at explanatory or predictive models. However, as will be discussed, the explanatory (or predictive) model must not be built using the same data as the exploratory model. Unfortunately, this distinction is not often recognized, and exploratory modeling is carried out. The "final" model, however, is treated as if it were a pre-specified model. In the context of exploratory models, usefulness means that testable hypotheses can be developed for future studies to investigate.

General comments

Regardless of the explicit modeling goal, it is essential to remember that all models represent an attempt at approximating some underlying reality. Hence, there are likely many equally useful approximations for any given situation. One of the (often quoted) guiding principles of modeling is called the "parsimony principle". In practice, it means that if a more complex model provides no meaningful advantage over a simpler model, where both models were pre-specified, the simpler model is preferred. However, it seems that much model building in the literature takes the view that this principle means that the analyst should attempt to find the simplest model that best explains a given data set. The flaw in this reasoning is that the goal of modeling is not to best explain the given data but to arrive at better undertanding of some population of interest.

Discussion of approach A

The approach of Researcher A, described above, is commonly used. Unfortunately, there are a host of problems associated with commonly applied techniques for variable and model selection. It is important to note that variable and model selection remains a topic of research.

Perhaps the biggest issue is that when the same data are used to formulate *and* fit a model, the usual linear regression theory does not apply.[1] In other words, the theory that lets us compute *P*-values and confidence intervals is no longer valid, meaning that *P*-values and confidence intervals are wrong. These problems extend

beyond linear models to generalized linear models such as logistic regression.

Other problems are that effect estimates are likely to be biased (i.e., too big), prediction intervals too narrow, and *P*-values too small.[2] Because these variable and model selection processes are intended to produce a *best* fit model for the data, it looks better than it really is (i.e., more optimistic), and diagnostics are unlikely to reveal serious problems because this model is *best*.[1] This optimism is due to overfitting and means that, if the best-fit model from a given set of data is fitted on a new sample from the same population, it will not fit as well. Hence, the best-fit model is not generalizable.

The single-variable pre-screening described here has an additional serious problem. It fails to identify residual confounding.[2] For example, suppose the male subjects were slightly younger than the female subjects, but more male subjects were smokers. The sex variable could fail to show up in pre-screening. However, after smoking status and age are accounted for simultaneously in the model, sex could become statistically significant. This type of situation is never found when using single-variable pre-screening.

Despite the serious problems with this approach and others like it, these approaches continue to be used with little opposition. This practice reveals two fallacies in statistical thinking. First, there is a belief that small *P*-values imply important effects, and large *P*-values imply unimportant effects. This reflects an incorrect understanding of *P*-values.[3] A practical expression of this fallacy is belief that only *statistically significant* variables are *allowed* to be in a model. The second fallacy in reasoning forgets that a model represents an attempt to explain population associations, whereas typical data-driven approaches explain associations present in a specific data set.

Discussion of approach B

The approach of Researcher B suffers none of the problems associated with approach A, although overfitting could occur if too many variables relative to the sample size are included.

Naturally, it takes a bit more thought and careful planning up front to arrive at a sensible conceptual model to fit to a data set. The process of creating a conceptual model should consider variables with a plausible theoretical reason to be important as well as variables consistently demonstrated as important in the literature. One must remember that the goal is to define a model that is useful for understanding the role of the primary exposure, not for determining the true underlying model (which is impossible anyway). For an explanatory model, this means focusing mainly on variables that are likely to be confounders or effect modifiers (i.e., "interactions" in statistical jargon). Identifying potential effect modifiers is more challenging than identifying potential confounders. These modifiers should be limited to those with a solid theoretical justification or that have been suggested in previous exploratory models, provided they are theoretically plausible.

Even so, it is quite easy to identify a sufficiently large number of variables so the sample size is insufficient to fit a model. Researchers sometimes use this fact as a justification for automated variable selection approaches, but it is not the correct approach. Arguably, if a data set is too small to fit a carefully defined model, a larger sample should be obtained (if possible). Randomized, controlled trials can get by on smaller sample sizes because confounding is removed by randomization so the primary, unadjusted comparison is valid. In contrast, observational studies require much larger sample sizes to adjust adequately for the relevant confounding variables. In other words, automated variable selection is not a substitute for a larger sample size. Even so, it cannot be said that valuable contributions cannot be made from smaller observational data sets. This situation may be viewed more realistically as providing exploratory insight rather than a robust explanatory (or predictive) model.

It is often puzzling to researchers when *known* confounders do not show any *statistically significant* associations in a fitted model. Conversely, automated variable selection does not guarantee that known confounders are retained in a model. There are a number of reasons why a known confounder may not show a statistically significant association, even for a variable that truly is a confounder in the population. The first possibility has to do with patient selection. Many observational studies (and nearly all trials) have some form of patient selection, which could result in variables having a reduced range (or variability) compared to those in the population. Variation in an explanatory variable is necessary (but not sufficient) to show an association with an outcome. Another possibility is that a model with that exact collection of variables has never before been fit. In multivariable regression models, it is generally impossible to anticipate how adjusting for a set of variables will affect another variable, especially in human subjects where variables are all interrelated. If a variable is truly unimportant, including it in a model does no harm other than using up some degrees of freedom, which should not be a problem if the data set was sufficiently large at the outset.

Table Definitions of some common terms in statistical modeling

Term	Definition
Outcome	This is the *response* variable of interest. In the example there are two outcomes mentioned; (1) occurrence of a perioperative complication and (2) hospital length of stay
Exposure	This is the variable whose relation to the outcome is of primary interest. In a randomized controlled trial, this would be treatment. In the example, the exposure is smoking status
Confounder	A confounder is a variable that is related to both the exposure and outcome and that, if not properly accounted for, can mask the true nature of the association between the exposure and the outcome. In the example, age, sex, etc. are potential confounders
Predictor	A predictor is a variable that is related only to an outcome and does not mask the association between the exposure and outcome. The variables available in the example could be predictors if they were not confounders
Effect modifier	This is a variable that results in different exposure effects depending on its value. For example, if female smokers experienced a greater risk of complications than male smokers, sex would be an effect modifier
Multicollinearity	Multicollinearity is a complicated phenomenon. It occurs when a collection of variables is related in such a way that one variable can be reasonably predicted in a linear manner (i.e., addition and subtraction of the other variables or multiples thereof) from the remainder. As a simple example, arm span is very close to the height for most people. Thus, you could predict height from arm span. Also, if you included both variables in a model, collinearity (multicollinarity when only two variables are involved) would be present. In extreme cases multicollinearity can prevent a model from being fit, but the general problem is that sorting out effects among a set of multicollinear variables is difficult. Thus, it is particularly a problem for explanatory models
Stepwise regression	For simplicity I grouped a collection of methods for automatic variable selection (forward selection, backward elimination, and the combination - usually labeled as stepwise). All of these methods share the characteristic that, at each step, a variable is either selected for addition to, or removal from, a model based on a *P*-value
Overfitting	This is a situation where a model is too complex for a data set. This results in a model appearing to be better fit than it would be if it were fit on a different data set (from the same population). Overfitting results in models that are not generalizable. Stepwise regression is a very complex model because each step needs to be accounted for, resulting in a huge number of fits comprising the model

Conclusion

I have attempted herein to describe some of the serious problems of the popular approach of Researcher A and address some of the common criticisms levelled at the preferred approach of Researcher B. I conclude by providing some suggestions for building explanatory models that may be useful.

1. For explanatory (or predictive) models, determine the variables you wish to collect (or are available) that are most likely to confound the relation of the primary exposure with the outcome. This decision can be based on theoretical grounds or from the literature. If using the literature, consider only variables that have repeatedly been shown to be important in good exploratory models.
   Exploratory models are likely to employ some form of data-dependent variable selection, although single-variable pre-screening probably should be avoided for the serious reasons mentioned earlier. Indeed, multiple approaches might be considered as an added measure of caution. Additionally, the choice of final candidate variables should be guided more by considering how the estimated effect sizes compare with clinically meaningful effects rather than *P*-values. In the end, a "final model" is postulated. The important point here is that such a final model should be viewed and reported only as a candidate for verification in a future study employing new data.
2. Do not categorize continuous variables and do not assume that continuous variables have a linear association with an outcome. For example, do not categorize age as < 70 *vs* ≥ 70 yr. Consider non-linear techniques, such as splines (a flexible technique for modelling arbitrary non-linear relations) or polynomials. That said, if the primary exposure variable of interest is continuous, there may be some value in categorization for the purposes of data summaries. Even so, it is preferable *not* to categorize in statistical models. Furthermore, when the continuous variable is used simply for adjustment, no benefit is realised by categorization. Finally, a good predictive model is likely too complex to apply without some aid. For bedside application, using a nomogram may be helpful.
3. It is important to determine how large a sample is needed to support a proposed model. Sample size determination for observational studies is more challenging than for randomized, controlled trials because a simple comparison is not the end goal. Sample sizes for robust explanatory (or predictive)

models generally need to be quite large. Harrell[2] provided some guidelines for the relation between the sample size and number of variables a model can support. If you are conducting a prospective (e.g., cohort study) study, pre-specifying the variables could assist with determining minimum sample sizes. Many observational studies, however, use existing data sets, so sample size is out of the investigator's control. In that case, these guidelines permit the researcher to force suitable limits on the modeling exercise. If an available sample is too small for explanatory modeling, exploratory models are indicated. Even in this case, though, one must take care not to overuse the data.

4. Effect modifiers (i.e., interactions) add an extra dimension of complexity to model interpretation and model selection. Therefore, as already mentioned, there should be very solid reasons for considering them. This is one of the situations when the "parsimony principle" is valuable. A useful approach here is to compare the model having no interactions *vs.* one having all of the proposed interactions. If the more complicated model is "better," keep it. Otherwise, choose the simpler model. In this instance, performing a single test avoids some of the issues encountered with multiple testing and limits the effects of overfitting if a sufficiently large significance level is employed.[2] For exploratory models, a formal comparison is less important than considering whether effect estimates are clinically meaningful.
5. Fit and interpret the pre-specified model. Do not worry so much about *P*-values but, rather, report the key effect estimates and confidence intervals (often 95%) of desired precision.
6. If there is interest in determining the variables that appear to be confounders, these investigations should use the full model rather than the classic approach of considering only the exposure and a single confounder. Specifically, once the full model is fit, one could fit a number of additional models where one of the potential confounders is removed, noting the main exposure effect. The full model is still the one to be reported. These additional models are used only to assess changes in the exposure effect as a result of ignoring a potential confounder while adjusting for other variables. The important question surrounding a potential confounder is whether it is a confounder after adjustment.
7. Finally, the usual model diagnostics should be performed. Internal validation using the bootstrap (an advanced statistical method using repeated resampling of existing data) should then be carried out to assess model optimism.[2]

This list may appear daunting in that a wider range of expertise is required than most researchers possess. Thus, a team approach is recommended wherein the expertise of two or more content experts is coupled with the expertise of a statistician familiar with these principles.

It is my hope that this editorial has exposed the weaknesses of commonly employed methods and provided helpful guidance for producing useful models.

Comment bâtir des modèles de régression pour des études observationnelles (et comment ne PAS le faire!)

Introduction

Tous modèles sont faux, mais certains sont utiles

Ce puissant aphorisme a été attribué sous différentes formes à l'éminent statisticien George Box. En dépit de son apparente simplicité, il reflète un certain nombre de problèmes plus profonds. Il nous rappelle tout d'abord que toutes les analyses statistiques sont, d'une certaine façon, un moyen de modéliser les phénomènes du monde réel. Ensuite, il sous-entend qu'aucun modèle statistique n'est parfait et que, dans un certain sens, ils doivent être faux. Les modèles peuvent être imparfaits pour différentes raisons, notamment des hypothèses erronées, une spécification incorrecte, des stratégies faussées de construction de modèles, etc. Enfin, cela nous rappelle que même si les modèles que nous fabriquons sont imparfaits, certains peuvent néanmoins être utiles. Bien que l'utilité d'un modèle soit étroitement liée à l'objectif principal d'une analyse, nous définirons dans la discussion qui suit un modèle *utile* comme étant un modèle qui nous informe sur la population avec une distorsion minimum et qui permet au chercheur de généraliser à partir de cet échantillon à l'ensemble de la population avec une confiance raisonnable. Un important corollaire est que certains modèles ne sont pas utiles. Personne ne sera surpris à l'idée qu'une bonne méthodologie est un ingrédient essentiel à la production de modèles utiles. Malheureusement, une bonne part de la construction de modèles rencontrés dans la littérature emploie une méthodologie biaisée. Le reste de cet article aborde certains de ces défauts et propose des alternatives. Pour les mettre en perspective, nous commencerons par un scénario hypothétique.

Deux chercheurs en anesthésie souhaitent étudier les rapports entre le tabagisme et, à la fois, les complications périopératoires et la durée de l'hospitalisation. Ils ont tous

deux accès aux résultats qui les intéressent via les dossiers électroniques des patients, leur statut par rapport au tabagisme, ainsi qu'à d'autres caractéristiques potentiellement importantes (par ex., l'âge, le sexe, le type d'anesthésie) pour des milliers de patients. Ils comptent utiliser une régression logistique pour modéliser l'existence d'une complication et une régression linéaire pour modéliser la durée d'hospitalisation.[A] Le chercheur A prévoit de commencer par une présélection monofactorielle. Cela signifie que chaque variable d'intérêt sera testée à la recherche d'une association avec le critère d'évaluation. Toutes les variables pour lesquelles $P < 0{,}05$ passent alors à la phase suivante. Au cours de la deuxième phase, toutes les variables ayant franchi l'étape de sélection initiale entreront dans un modèle et une élimination régressive sera utilisée pour parvenir à l'ensemble définitif de variables *statistiquement significatives*. Suivant une démarche totalement différente, le chercheur B commence par identifier, à partir de la documentation disponible ou de bases théoriques, des variables connues (ou qui sont fortement suspectées) d'avoir un rapport avec le statut tabagique et le critère d'évaluation. Ensuite, toutes les variables identifiées ont été incluses dans le modèle et conservées sans tenir compte de la valeur de *P*.

Devant ces deux démarches opposées, il est probable qu'en dépit de la popularité de la démarche A dans la littérature, de nombreux chercheurs soupçonneront sérieusement que la démarche B aura plus de chances de déboucher sur un modèle utile (au sens défini plus haut). En vérité, la démarche A (et d'autres semblables) doit être évitée et la démarche B devrait être la méthode préférée pour l'analyse des données. La suite de cet éditorial étudiera les défauts de la démarche A et fournira des conseils sur l'utilisation de la démarche B afin d'obtenir des modèles utiles.

Objectifs des modèles

Avant de procéder à l'analyse des deux démarches présentées, il est utile de revoir brièvement deux objectifs des modèles statistiques : prédiction contre explication. Même s'il y a une certaine communauté de vues entre ces deux démarches, certains problèmes doivent être abordés avec circonspection.

Modèles explicatifs

Le but des modèles explicatifs est de comprendre. C'est-à-dire que l'on souhaite comprendre et quantifier les effets des variables sur un critère d'évaluation. Vous voulez donc que les estimations de l'effet soient (idéalement) libres de tout biais. Il vous faut prendre en compte des variables confondantes d'une relation (c'est-à-dire, qui les masquent) et veiller à se protéger d'une multicolinéarité, mais les variables qui *prédisent* uniquement les critères d'évaluation sont moins importantes à prendre en compte (voir l'explication de ces termes statistiques et d'autres dans le Tableau). Le modèle explicatif est probablement le modèle le mieux adapté et le plus courant pour les données observationnelles. Comme l'objectif de notre scénario hypothétique est de comprendre comment le tabagisme est relié à certains critères d'évaluation, cela correspond au paradigme explicatif.

Modèles prédictifs

L'objectif des modèles prédictifs est de prédire *avec exactitude* un critère d'évaluation. C'est-à-dire que le modèle doit être capable de bien prédire les réponses dans la population. Dans ce type de modèle, les éléments prédictifs sont importants, mais il n'est pas essentiel d'aller jusqu'aux explications des associations de variables. Même la multicolinéarité peut ne pas être un gros problème, sous réserve que le modèle ait de bonnes propriétés prédictives. Il est important de réaliser que le modèle qui s'adapte *le mieux* aux données n'est vraisemblablement pas le meilleur modèle prédictif, car un surapprentissage (voir le Tableau) est susceptible d'exister dans le *meilleur* modèle. Si, dans ce scénario hypothétique, les chercheurs essayaient de parvenir à prédire avec fiabilité quels patients développeraient des complications postopératoires en fonction de leurs caractéristiques avant l'intervention, il leur faudrait élaborer un modèle prédictif. Nous ne développerons pas davantage ici les modèles prédictifs. Mais il est intéressant de noter qu'un modèle prédictif utile est un modèle qui prédit « avec exactitude » l'évolution des futurs patients. Il se préoccupe moins d'obtenir des explications utiles sur les rapports sous-jacents, pour autant que la prédiction soit bonne.

Modèles exploratoires

Il peut arriver que l'on sache relativement peu de choses sur un problème particulier. Dans de telles situations, les modèles peuvent servir à établir des hypothèses sur les rapports potentiels qu'il est nécessaire de tester au cours d'études ultérieures. Si les chercheurs A et B ne s'intéressaient qu'à savoir quels facteurs pouvaient être en rapport avec les résultats d'intérêt, des modèles exploratoires auraient été indiqués. Il y a en fait une réflexion exploratrice intriquée dans la démarche du chercheur A. L'objectif des modèles exploratoires peut être de parvenir éventuellement à des modèles explicatifs ou prédictifs. Cependant, comme nous allons le voir, le

Tableau Définitions de quelques termes courants utilisés pour la modélisation statistique

Terme	Définition
Critère d'évaluation	Il s'agit de la *réponse* pour la variable d'intérêt. Dans l'exemple, deux critères d'évaluation sont mentionnés; (1) survenue d'une complication périopératoire (2) durée de l'hospitalisation
Exposition	Il s'agit de la variable dont la relation avec le critère d'évaluation est d'intérêt primordial. Dans un essai randomisé contrôlé, il pourrait s'agir du traitement. Dans l'exemple, l'exposition est le statut envers le tabagisme
Élément confondant	Un élément confondant est une variable qui est liée à la fois à l'exposition et au critère d'évaluation et qui, s'il n'est pas correctement pris en compte, peut masquer la véritable nature de l'association entre l'exposition et le critère d'évaluation. Dans l'exemple, l'âge, le sexe, etc., sont des éléments confondants potentiels
Élément prédictif	Un élément prédictif est une variable qui n'est liée qu'à un critère d'évaluation et qui ne masque pas l'association entre l'exposition et ce critère. Les variables disponibles dans l'exemple pourraient être des éléments prédictifs s'il ne s'agissait pas d'éléments confondants
Modificateur de l'effet	Il s'agit d'une variable qui aboutit à différents effets de l'exposition en fonction de sa valeur. Si, par exemple, des fumeuses ont présenté un plus grand risque de complications que les hommes qui fument, le sexe serait un modificateur de l'effet
Multicolinéarité	La multicolinéarité est un phénomène compliqué. Elle survient quand une série de variables est liée de telle façon qu'il est raisonnablement possible de prédire une variable de façon linéaire (c'est-à-dire, l'addition ou la soustraction des autres variables ou de leurs multiples) du reste. Un exemple simple est l'envergure des bras qui est très proche de la taille chez la plupart des individus. Vous pourriez ainsi prédire la taille d'une personne à partir de son envergure. Si vous incluez les deux variables dans un modèle, une colinéarité (multicolinéarité quand seulement deux variables sont impliquées) pourrait être présente. Dans les cas extrêmes, une multicolinéarité peut empêcher l'adaptation d'un modèle, mais le problème général est que le tri des effets dans un ensemble de variables multicolinéaires est difficile. C'est donc un problème particulier pour les modèles explicatifs
Régression séquentielle (ou pas-à-pas)	Dans un but de simplification, j'ai regroupé une série de méthodes pour la sélection automatique des variables (sélection ascendante, élimination descendante, et leur combinaison - habituellement qualifiée de pas-à-pas). Toutes ces méthodes ont une caractéristique en commun : à chaque étape, une variable est soit sélectionnée pour être ajoutée ou retirée d'un modèle reposant sur la valeur de P
Surapprentissage	Il s'agit d'une situation où un modèle est trop complexe pour un ensemble de données. Cela aboutit à un modèle qui semble mieux adapté qu'il ne le serait s'il était adapté à un ensemble de données différent (provenant de la même population). Le surapprentissage aboutit à des modèles qui ne peuvent être généralisés. La régression séquentielle est un modèle très complexe parce que chaque étape doit être prise en compte, aboutissant à un nombre gigantesque d'adaptations comprises dans le modèle

modèle explicatif (ou prédictif) ne doit pas être bâti à partir des mêmes données que le modèle exploratoire. Malheureusement, cette distinction n'est pas toujours reconnue et le modèle exploratoire est réalisé. Le modèle « final » est traité comme s'il s'agissait d'un modèle préspécifié. Dans le contexte des modèles exploratoires, le caractère utilisable signifie que les hypothèses qu'on peut tester peuvent être élaborées afin d'être soumises à des études ultérieures.

Commentaires généraux

Indépendamment de l'objectif explicite de la création du modèle, il est essentiel de se rappeler que tous les modèles représentent une tentative de représentation d'une réalité sous-jacente. En conséquence, il y a vraisemblablement de nombreuses approximations tout aussi utiles les unes que les autres pour chaque situation donnée. Un des principes d'orientation (souvent cité) pour la création de modèles est appelé le « principe de parcimonie ». Cela signifie qu'en pratique, si un modèle plus complexe ne procure pas d'avantages significatifs par rapport à un modèle plus simple lorsque les deux modèles ont été prédéterminés, on préférera le modèle le plus simple. Mais il semble à l'examen de la littérature qu'un grand nombre de publications sur la construction de modèles considèrent que ce principe signifie que l'analyste doit tenter de trouver le modèle le plus simple expliquant au mieux un ensemble de données spécifiques. Le défaut de ce raisonnement tient à ce que l'objectif de la construction de ce modèle n'est pas d'expliquer au mieux des données particulières, mais de parvenir à mieux connaître une partie de la population d'intérêt.

Discussion sur la démarche A

La démarche du chercheur A, décrite plus haut, est couramment utilisée. Il y a malheureusement un grand nombre de problèmes associés aux techniques habituellement utilisées pour la sélection des variables et du modèle. Il est important de noter que la sélection des variables et du modèle reste un sujet de recherche.

Le plus grand problème est peut-être que lorsque les mêmes données sont utilisées pour formuler un modèle *et* s'y adapter, la théorie habituelle de la régression linéaire ne

s'applique pas.[1] En d'autres termes, la théorie qui nous laisse calculer les valeurs de *P* et les intervalles de confiance n'est plus valide, ce qui signifie que ces derniers sont faux. Ces problèmes vont au-delà des modèles linéaires et atteignent les modèles linéaires généralisés, tels que la régression logistique.

Parmi les autres problèmes, les estimations de l'effet risquent d'être faussés (c'est-à-dire, trop importantes), les intervalles de prédiction trop étroits et les valeurs *P* trop petites.[2] Dans la mesure où ces processus de sélection des variables et modèles sont conçus pour produire le modèle qui correspond *le mieux* aux données, ils paraissent meilleurs qu'ils ne le sont en réalité (c'est-à-dire plus optimistes) et les diagnostics sont peu susceptibles de révéler des problèmes graves parce que ce modèle est *le meilleur*.[1] Cet optimisme est dû au surapprentissage et signifie que si le modèle le mieux adapté d'un ensemble de données spécifiques est appliqué à un nouvel échantillon de la même population, il ne s'adaptera pas aussi bien. Par conséquent, le modèle le mieux adapté n'est pas généralisable.

La présélection d'une variable unique décrite ici présente un autre sérieux problème. Elle ne parvient pas à identifier les facteurs résiduels confondants.[2] Supposons par exemples que les hommes étaient légèrement plus jeunes que les femmes, mais qu'il y avait davantage de fumeurs parmi les hommes. La variable « sexe » pourrait ne pas apparaître dans la présélection. Cependant, une fois le tabagisme et l'âge pris simultanément en compte dans le modèle, le sexe pourrait devenir statistiquement significatif. Ce type de situation n'est jamais trouvé avec l'utilisation d'une présélection avec variable unique.

En dépit des problèmes graves, et d'autres semblables, soulevés par cette démarche, cette dernière continue d'être utilisée sans grande opposition, une pratique qui révèle deux faux raisonnements de la pensée statistique. Premièrement, la croyance veut qu'une valeur faible de *P* implique des effets importants et qu'au contraire, des valeurs élevées de *P* impliquent des effets sans importance. Cela reflète une mauvaise interprétation des valeurs de *P*.[3] D'un point de vue pratique, l'expression de cette erreur est la croyance que seules les variables *statistiquement significatives* sont *autorisées* à entrer dans un modèle. La seconde erreur de raisonnement oublie qu'un modèle représente une tentative d'explication d'associations de population alors que les démarches habituelles reposant sur les données expliquent les associations présentes dans un ensemble de données spécifiques.

Discussion sur la démarche B

La démarche du chercheur B ne souffre d'aucun des problèmes associés à la démarche A, bien qu'un surapprentissage puisse survenir si de trop nombreuses variables en rapport avec la taille de l'échantillon sont incluses.

Naturellement, cela demande davantage de réflexion et de planification méticuleuse avant de parvenir à un modèle conceptuel valide qui s'adapte à un ensemble de données. Le processus de création d'un modèle conceptuel doit envisager les variables ayant une raison théorique plausible d'avoir de l'importance ainsi que des variables dont l'importance a été constamment démontrée dans la littérature. Il faut se souvenir que l'objectif est de définir un modèle qui soit utile pour la compréhension du rôle de l'exposition principale, pour la détermination du modèle sous-jacent véritable (ce qui est de toute manière impossible). Pour un modèle explicatif, cela veut dire se concentrer principalement sur des variables susceptibles d'être des facteurs confondants ou des modificateurs de l'effet (en d'autres termes, des « interactions » dans le jargon statistique). L'identification d'éventuels modificateurs de l'effet est plus difficile que l'identification d'éventuels facteurs confondants. Ces modificateurs doivent être limités à ceux disposant d'une solide justification théorique ou à ceux ayant été déjà proposés dans des modèles explicatifs antérieurs sous réserve qu'ils soient plausibles d'un point de vue théorique.

Même dans ce cas, il peut être assez facile d'identifier un nombre suffisamment grand de variables de sorte que la taille de l'échantillon sera insuffisante pour correspondre au modèle. Les chercheurs utilisent parfois ce fait comme justification des démarches de sélection automatisée des variables, mais cela n'est pas la bonne démarche. D'un autre côté, si un ensemble de données est trop petit pour s'adapter à un modèle défini avec soin, il faut obtenir un plus grand échantillon (dans la mesure du possible). Les essais randomisés, contrôlés, peuvent utiliser des échantillons plus petits parce que la confusion est supprimée par la randomisation, de sorte que la comparaison principale, non ajustée, est valide. En revanche, les études observationnelles nécessitent des échantillons beaucoup plus grands pour s'ajuster correctement aux variables confondantes pertinentes. En d'autres termes, la sélection automatisée des variables ne remplace pas une plus grande taille de l'échantillon. Même dans ce cas, on ne peut pas dire que des plus petits ensembles de données observationnelles ne permettront pas de tirer des contributions valables. Cette situation peut être perçue de façon plus réaliste comme procurant un point de vue exploratoire plutôt qu'un solide modèle explicatif (ou prédictif).

Les chercheurs sont souvent désarçonnés quand des éléments confondants *connus* ne montrent aucune association *statistiquement significative* dans un modèle adapté. En revanche, la sélection automatisée des variables

ne garantit pas que des éléments confondants connus seront retenus dans un modèle. Il y a un certain nombre de raisons pour lesquelles un élément confondant connu peut ne pas afficher d'association statistiquement significative, même pour une variable qui est véritablement confondante dans la population. La première possibilité tient à la sélection des patients. De nombreuses études observationnelles (et presque tous les essais) incluent une forme quelconque de sélection qui pourrait déboucher sur des variables ayant une plage (ou variabilité) réduite, par rapport à celles de la population. Une variation d'une variable explicative est nécessaire (mais insuffisante) pour montrer une association avec un critère d'évaluation. Une autre possibilité est qu'un modèle ayant une collection exacte de variables n'a jamais été adapté antérieurement. Dans les modèles de régression multifactorielle, il est habituellement impossible d'anticiper comment l'ajustement pour un ensemble de variables va affecter une autre variable, en particulier chez des sujets humains dont toutes les variables sont interdépendantes. Si une variable est vraiment sans importance, l'inclure dans un modèle ne fait aucun autre tort que d'utiliser quelques degrés de liberté, ce qui ne devrait pas être un problème si l'ensemble de données était suffisamment grand d'entrée de jeu.

Conclusion

J'ai tenté de décrire ici quelques problèmes graves liés à la démarche populaire du chercheur A et répondre à quelques-unes des critiques fréquentes adressées à la démarche préférée du chercheur B. Je conclus en proposant quelques suggestions qui pourraient être utiles pour la construction de modèles explicatifs.

1. Pour les modèles explicatifs (ou prédictifs), définissez les variables que vous souhaitez (ou êtes en mesure de) collecter et qui sont le plus susceptibles de brouiller le rapport existant entre l'exposition principale et le critère d'évaluation. Cette décision peut s'appuyer sur une base théorique ou sur la documentation existante. En cas d'utilisation des publications, n'envisager que les variables qui se sont avérées importantes de façon répétée dans de bons modèles exploratoires.
   Les modèles exploratoires sont susceptibles d'utiliser une certaine forme de sélection des variables dépendantes de données, bien qu'une présélection de variable unique doive être évitée pour les importantes raisons mentionnées précédemment. En fait, de nombreuses démarches pourraient être envisagées comme mesure supplémentaire de précaution. De plus, la façon dont on peut comparer les tailles estimées de l'effet avec des effets cliniquement probants devrait guider davantage le choix final des variables candidates, de préférence aux valeurs de *P*. Un « modèle définitif » peut être enfin proposé. Le point important ici est que ce modèle définitif doit être considéré et décrit seulement comme un candidat à vérifier au cours d'une étude future utilisant de nouvelles données.
2. Ne classez pas les variables continues en catégories et ne supposez pas que ces variables continues sont associées de manière linéaire à un résultat. Il ne faut pas, par exemple, créer des groupes d'âge tels que < 70 ans contre ≥ 70 ans. Il faut envisager des techniques non linéaires, telles que les splines (une technique souple pour la modélisation de relations arbitraires non linéaires) ou des fonctions polynomiales. Cela dit, si la principale variable d'exposition d'intérêt est continue, il pourrait y avoir un certain avantage à constituer des catégories pour permettre une synthèse des données. Mais même dans ce cas, il est préférable de ne *pas* créer de catégories dans des modèles statistiques. En outre, quand une variable continue est utilisée simplement pour un ajustement, la catégorisation ne procure aucun bénéfice. Enfin, l'application d'un bon modèle prédictif est probablement trop compliquée sans un certain degré d'assistance. L'utilisation d'un nomogramme peut être utile au chevet des patients.
3. Il est important de déterminer quelle doit être la taille d'un échantillon pour soutenir un modèle proposé. La détermination de la taille de l'échantillon pour les études observationnelles est plus difficile que pour les études randomisées et contrôlées, car une simple comparaison n'est pas l'objectif final. La taille des échantillons pour de solides modèles explicatifs (ou prédictifs) doit habituellement être assez importante. Harrell[2] a fourni quelques lignes directrices sur la relation entre la taille d'un échantillon et le nombre de variables qu'un modèle peut supporter. Si vous menez une étude prospective (une étude de cohorte, par exemple), la prédétermination des variables pourrait aider à établir la taille minimum des échantillons. Cependant, de nombreuses études observationnelles utilisent des ensembles existants de données, si bien que l'investigateur ne peut contrôler la taille de l'échantillon. Dans ce cas, ces lignes directrices permettent au chercheur d'imposer des limites convenables à l'exercice de modélisation. Si un échantillon disponible est trop petit pour une modélisation explicative, des modèles exploratoires sont indiqués. Mais, même dans ce cas, il faut veiller à ne pas sur-utiliser les données.
4. Les modificateurs de l'effet (c'est-à-dire, les interactions) ajoutent une dimension supplémentaire

à la complexité de l'interprétation et de la sélection du modèle. Donc, comme déjà dit, il doit y avoir de très bonnes raisons pour les envisager. C'est là une des situations dans lesquelles le « principe de parcimonie » montre tout son intérêt. Une démarche utile consiste alors à comparer le modèle n'ayant pas d'interactions avec un modèle ayant toutes les interactions proposées. Si le modèle le plus compliqué est « meilleur », gardez-le. Sinon, choisissez le modèle plus simple. Dans cet exemple, la réalisation d'un seul test évite quelques-uns des problèmes rencontrés avec des tests multiples et limite les effets du surapprentissage si on utilise un niveau de signification suffisamment grand.[2] Pour les modèles exploratoires, une comparaison formelle est moins importante que de se demander si les estimations de l'effet sont cliniquement probantes.

5. Adapter et interpréter le modèle préspécifié. Ne vous souciez pas tant des valeurs de *P*, mais plutôt, rapportez les estimations de l'effet essentiel et les intervalles de confiance (souvent de 95 %) de la précision désirée.
6. S'il y a intérêt à déterminer les variables qui paraissent être confondantes, ces recherches doivent utiliser le modèle complet plutôt que la démarche classique consistant à ne considérer que l'exposition et un seul élément confondant. Plus spécifiquement, une fois le modèle adapté, on pourrait adapter un certain nombre de modèles supplémentaires dans lesquels on retire l'un des éléments confondants potentiels, en notant l'effet de l'exposition principale. Le modèle complet est encore le modèle à rapporter. Ces modèles supplémentaires ne sont utilisés que pour évaluer des changements dans l'effet de l'exposition résultant de l'ignorance d'un élément confondant potentiel, tout en faisant des ajustements pour les autres variables. La question importante entourant un élément confondant potentiel est de savoir si c'est effectivement un élément confondant après l'ajustement.
7. Enfin, les tests diagnostiques usuels doivent être appliqués au modèle. Une validation interne utilisant le processus d'amorce ou bootstrap (une méthode statistique avancée faisant appel à un rééchantillonnage répété de données existantes) doit être réalisée pour évaluer l'optimisme du modèle.[2]

Cette liste peut paraître décourageante dans la mesure où une expertise plus étendue que celle dont disposent la majorité des chercheurs est nécessaire. Aussi, un travail d'équipe est recommandé, associant l'expertise de deux experts de contenu, ou plus, à celle d'un statisticien ayant l'habitude de ces principes.

J'ai l'espoir que cet éditorial a permis d'exposer les faiblesses des méthodes couramment employées et a fourni un guide intéressant pour la production de modèles utiles.

Conflicts of interest None declared.

Editorial responsibility This submission was handled by Dr. Philip M. Jones, Associate Editor, *Canadian Journal of Anesthesia.*

Conflits d'intérêts Aucun déclaré.

Responsabilité éditoriale Cet article a été traité par le Dr Philip M. Jones, rédacteur adjoint, *Journal canadien d'anesthésie.*

References

1. *Chatfield C*. Model uncertainty, data mining and statistical inference. J R Statist Soc A 1995; 158: 419-66.
2. *Harrell FE*. Regression Modeling Strategies. With Applications to Linear Models, Logistic Regression, and Survival Analysis. NY: Springer; 2001.
3. *Wasserstein RL*, *Lazar NA*. The ASA's Statement on p-Values: Context. Process and Purpose. The American Statistician 2016; 70: 129-33.