What You See May Not Be What You Get: A Brief, Nontechnical Introduction to Overfitting in Regression-Type Models

MICHAEL A. BABYAK, PHD

Objective: Statistical models, such as linear or logistic regression or survival analysis, are frequently used as a means to answer scientific questions in psychosomatic research. Many who use these techniques, however, apparently fail to appreciate fully the problem of overfitting, ie, capitalizing on the idiosyncrasies of the sample at hand. Overfitted models will fail to replicate in future samples, thus creating considerable uncertainty about the scientific merit of the finding. The present article is a nontechnical discussion of the concept of overfitting and is intended to be accessible to readers with varying levels of statistical expertise. The notion of overfitting is presented in terms of asking too much from the available data. Given a certain number of observations in a data set, there is an upper limit to the complexity of the model that can be derived with any acceptable degree of uncertainty. Complexity arises as a function of the number of degrees of freedom expended (the number of predictors including complex terms such as interactions and nonlinear terms) against the same data set during any stage of the data analysis. Theoretical and empirical evidence—with a special focus on the results of computer simulation studies—is presented to demonstrate the practical consequences of overfitting with respect to scientific inference. Three common practices—automated variable selection, pretesting of candidate predictors, and dichotomization of continuous variables—are shown to pose a considerable risk for spurious findings in models. The dilemma between overfitting and exploring candidate confounders is also discussed. Alternative means of guarding against overfitting are discussed, including variable aggregation and the fixing of coefficients a priori. Techniques that account and correct for complexity, including shrinkage and penalization, also are introduced. **Key words:** statistical models, regression, simulation, dichotomization, overfitting.

ANOVA = analysis of variance.

INTRODUCTION

In science, we seek to balance curiosity with skepticism. On one hand, to make discoveries and advance our knowledge, we must imagine and consider novel findings and ideas of all kinds. In the end, however, we also must subject those results to stringent tests, such as replication, to make sure that chance has not fooled us yet again (1). Modern data analytic methods such as multivariable regression models reflect these opposing processes quite well. We build models that we hope or imagine will reveal some significant scientific truth, but ultimately, because they are derived from the imperfect process of sampling, we must determine which of the "significant" findings we should believe and which we probably should not. The present article is a brief introduction to some concepts that can help us in this pursuit as it applies to regression-type modeling.

Most outside the community of statisticians are probably unaware that there has been something of a revolution in data analysis in the past 10 or so years. Modern computational power has not only made it easier to solve complex and large analytic problems but also allowed us to study, through a technique called *simulation*, the very act of collecting data and performing analyses. Through computer simulation studies (sometimes referred to as *Monte Carlo studies*), researchers with even modest personal computers can now study the performance of both new and traditional data-analytic techniques under a variety of circumstances. These simulation studies have taught us a great deal about the scientific merit of some of our conventions in data analysis and also have pointed toward new directions that may improve our practice as researchers.

In the present article, I will discuss a relatively narrow but important concept that has been considerably illuminated by simulation studies: the problem of capitalizing on the idiosyncratic characteristics of the sample at hand, also known as *overfitting*, in regression-type models. Overfitting yields overly optimistic model results: "findings" that appear in an overfitted model don't really exist in the population and hence will not replicate. Based at least in part on the simulation evidence, I hope to show that our inattention to the problem of overfitting may be hindering our scientific progress, and that the problem could be readily improved by avoiding some common mistakes and by adopting some alternative or additional steps in the process of model building. The discussion that follows is relevant to just about any attempt to model data, whether it be from observational studies or well-controlled randomized trials and experiments.

The ensuing text is taken from a workshop I recently presented to the members of the American Psychosomatic Society (2) and was intended for a nontechnical audience. I have purposely retained this nontechnical flavor and the informal language here so that these often esoteric ideas may be more accessible to a wider audience. Virtually none of the ideas I present here are my own. I am reporting the work of many others, explaining them in a way that I hope is understandable to readers of varying backgrounds and levels of expertise in data analysis. Specifically, the vast majority of this article is based on a relatively new statistical text by Harrell (3), who was responsible for my initial exposure to these ideas, and who continues to be a source of clarification and encouragement in my own pursuit of understanding these issues. Virtually all of the ideas presented here are discussed in much greater detail in the Harrell text (3).

As a final note of preface, the reader will soon find that I strongly endorse an approach in which as many of the ele-

From Duke University Medical Center, Durham, NC.

Address correspondence and reprint requests to Michael A. Babyak, PhD, Department of Psychiatry and Behavioral Science, Duke University Medical Center, Box 3119, Durham, NC 27710. E-mail: michael.babyak@duke.edu

Received for publication October 28, 2003; revision received February 19, 2004.

0033-3174/04/6603-0411

ments of the statistical model (such as the predictors that will be included) are specified a priori, with the additional condition that the model must not ask too much from the data at hand (a point that will be discussed in more detail later). I hope to show that the information provided by such models will generally constitute stronger scientific evidence than models that were achieved in other ways, such as univariate prescreening of predictors, or pulling variables in and out of a model to see which produce the best fit. I want to emphasize, however, that I am not at all opposed to going beyond those a priori analyses to explore the data. On the contrary, given the considerable personal and public resources we often expend in collecting and analyzing data, I believe we have a scientific and ethical obligation to extract as much information as we can from that data. What I am arguing, however, is that the various modeling techniques that we have at our avail fall roughly along a continuum in terms of the confidence we can ascribe to them. Predetermined models that have enough data to support them reside on 1 end of the continuum because they are more likely to produce replicable results, and on the other end reside the more exploratory approaches, such as graphing, fitting and refitting models, tinkering with assumptions, and so forth. Both represent important components of the scientific endeavor, but only if we make a frank assessment and description of their limitations.

PRELIMINARIES

Regression Models in Psychosomatic Research

The modern psychosomatic research literature is filled with reports of multivariable[1] regression-type models, most commonly multiple linear regression, logistic, and survival models. Although each has different underlying mathematical underpinnings, they share a general form that should be familiar to most, usually something like the following:

$$\text{response} = \text{weight}_1 \times \text{predictor}_1 + \text{weight}_2 \times \text{predictor}_2 + \ldots \text{weight}_k \times \text{predictor}_k + \text{error}$$

In other words, we generally are interested in finding a weighted combination of some set of variables that reproduces or *predicts* as well as possible the values that we have observed on the response or outcome variable. Typically, we evaluate the results of the model by conducting significance tests on the individual predictors and also by looking at how well the weighted combination of those variables predicts the response values. If we are good scientists, we also strive to test the model against a new set of data, collected under different circumstances, to assess how well the results generalize or replicate. Underlying this entire pursuit, of course, is the somewhat metaphysical assumption that somewhere out there in the population there is a true model, and that we can make a good approximation of that true model using only a portion, or sample, from that population. Thus, one overarching aim of modeling is to use the sample to come up with the correct set of predictors along with their weights, thus recovering the true model characteristics. If this aim is achieved, the model we develop will predict well not only in the sample data set at hand but also in new data sets. More broadly, if we have come close enough to identifying the true model (which, of course will still be only a crude approximation of the phenomenon we are studying), the science can move forward because we can then be pretty confident that the model is good enough to guide further research, clinical decisions, and policy. Finally, although there are some important distinctions between conducting models for the purpose of hypothesis testing per se versus modeling to make predictions, the fundamental principles of good model building are applicable to both aims.

Simulation Studies and the Advance of Data Analysis as a Science

As I noted in the opening, much of the remarkable acceleration in data analysis has been a direct consequence of improved computing power, which has allowed us to systematically study—via simulation—the very act of collecting and analyzing data. Simulation studies are like being able to study the accuracy of a diagnostic medical test under conditions in which the true diagnosis is already known. A statistical simulation study of modeling begins with a computer-generated population for which, much like knowing the correct diagnosis, the correct model is already known. The computer algorithm then simulates the activity of drawing a sample from the known population and conducting a regression model on the data from the sample. Because it is all performed on a computer, however, this act is repeated many thousands of times in a few seconds or minutes, each time using a newly drawn sample from the population (simulation studies often use 10,000 or more samples). The results from the many thousands of models are tallied and compared with the true population model. Most importantly, we can systematically manipulate various aspects of sampling and analytic activity. For example, we could use simulation to study how well an ordinary linear regression model recovers some known regression weight when the sample size is 100. To do this, we create a simulated population in which the true relation between x and y can be represented as, say, the equation $y = 0.4x$ and error. The simulation program draws a large number of random samples, in this case, 10,000 draws, each of $N = 100$, from that predefined population and then performs the regression analysis on each sample. The regression coefficients, or *weights*, from each sample's regression model are then compiled and described (Figure 1), empirically answering the question of how often we get a regression weight that is close to the known population value under the condition that $N = 100$.

Typically, simulation studies are designed like a factorial experiment, systematically manipulating various aspects of

[1]The term *multivariable* refers to the idea that more than 1 predictor, or *covariable*, is used in the same model. We often informally refer to these models as *multivariate*, but strictly speaking, the term *multivariate* is intended to describe models in which there is more than 1 response or outcome variable.

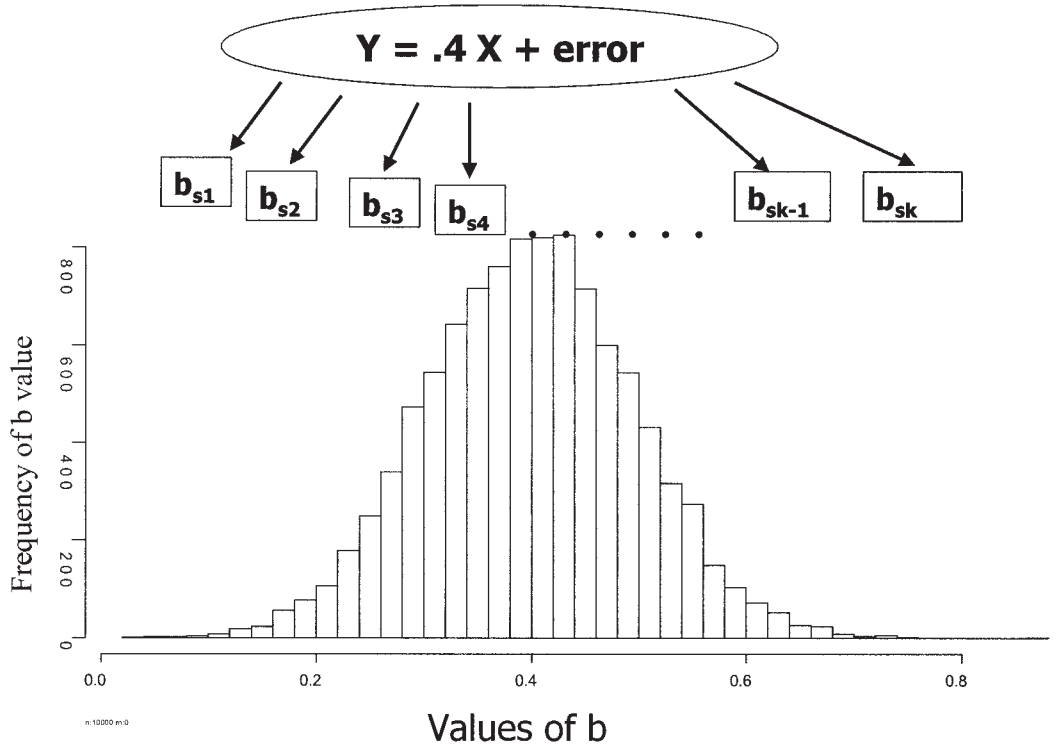

Figure 1. Example of a simple simulation study. A simulated population was created in which the equation y = 0.4x + error was true. Ten thousand random samples of $N = 100$ were drawn, and an ordinary least squares regression model, specified as y = bx + error, was estimated for each sample. The regression coefficient b was collected from each of the 10,000 models and plotted here by frequency. The location and shape of such a distribution can be examined to see whether it has the properties we would expect given our model assumptions.

sampling and analysis, such the sample size, the shape of the distribution of x or y, or the presence of missing data or noise variables, to name just a few conditions. Simulation studies have been a boon to the evolution of data analysis: they allow us to study whether new or even tried-and-true analytic methods meet the aim of capturing population values.

REQUIREMENTS FOR A GOOD MODEL

Right Model Form to Fit the Probability Question

I will jump ahead here and beg the question of whether modeling is a more useful approach than using specific tests, for example, using a regression model to test the group difference on a continuous response variable rather than performing a *t* test. Testing is not incorrect and sometimes can be sufficient to address the question at hand; models, however, can offer the same information as the tests and much more about the phenomenon under study, including an estimate of the size and direction of the effect, a means by which predictions can be made in new samples, and an estimate of the uncertainty of the result with respect to the observed data. As I proceed, however, it should be remembered that the issue of overfitting is equally relevant to statistical tests as it is to modeling, although perhaps not in as obvious a fashion. An exposition of each of the available models is beyond the scope of this article, but generally speaking, the choice of the model is determined by the nature of the dependent variable (and, more broadly, of course, the research question and the underlying probability model assumed). The vast majority of models that we use in psychological and medical research is subsumed under the generalized linear model (4). The generalized linear model allows a variety of distributions in the response variable, including normal, Poisson, binomial, negative binomial, multinomial, and gamma. Models with normally distributed responses are a special case of the generalized linear model and can be handled using the familiar general linear model forms, such as linear regression, analysis of variance (ANOVA), and analysis of covariance. Similarly, models with binary responses also are a special case of the generalized linear model, being synonymous with logistic regression, or in the case of ordered categories, the ordinal logistic model. Censored responses most typically require models in the time-to-failure family, such as the Cox regression model (5). For repeated or clustered response measures (eg, a multisite study, twin or couples data, or repeated measures), mixed or hierarchical linear models can be used for normally distributed responses, whereas there are extensions available for repeated or clustered nonnormal responses (6). As an aside, as a reviewer, I frequently see manuscripts that express special concern when the predictors in a model are not normally distributed, and hence make transformations or categories to overcome this problem. The shape of the distribution, or even the measurement form (ie, categorical vs. quantitative) of the predictors in a regression-type model, however, generally has no impact on the model except in some very

special instances. Consequently, under most circumstances, there is no real a priori need to transform or categorize a predictor on the basis of its distribution, or to be concerned that the predictor side has a combination of categorical and quantitative variables. As noted, it is the distribution of the response variable (or, more correctly, the errors or residuals that result from the prediction of the response) that matters.

Sample Size for Models and Overfitting

Assuming we have selected a reasonably representative sample from the population and have used reliable and valid measures of the variables of interest, we can use 1 of these models to make a guess at what the true model, or at least part of it, might look like. To achieve a model that will replicate in a new sample, we need to have an adequate sample size to generate reasonably accurate estimates of the unknowns (eg, b-weights in multiple regression). If we try to estimate too many unknowns—or, more technically speaking, if we use up too many degrees of freedom[2]—with respect to the number of observations, we will end up including predictor variables, or finding complicated relations (interactions, nonlinear effects) between the predictors and the response that indeed exist in the sample, but not in the population. In most circumstances, we also will reduce the power to detect true relations. Taken to its extreme, if the number of unknowns in a model is equal to the number of observations, the model will always fit the sample data perfectly, even if all the predictors are *noise*, ie, entirely unrelated to the response variable. Why does this happen? To take a very simple case, we all know that when we estimate a sample mean, its standard error (the probable fluctuation of the estimated mean over repeated samples of the same size) will be larger in small samples compared with larger samples. That is, the estimate of the mean will vary more widely over repeated samples when the sample size is smaller compared with when the sample size is relatively large. Similarly, regression-type models that use up too many degrees of freedom for the available sample size tend to produce weights that fluctuate considerably over repeated samples. The wider fluctuation over samples increases the chance of some of the regression weights being very large in a given sample, thus leading to an overly optimistic fit. Cast in slightly different light, statisticians often point out that estimating a regression with 10 predictors and 20 observations is in a sense the equivalent of estimating 10 separate 1-predictor regressions, each with a sample size of $N = 2$. The problem of instability of the regression coefficients is compounded by our natural tendency to focus on those larger regression weights (or worse yet, to cherry-pick the bigger or more significant weights for a final model), rather than choosing a priori which ones we will pay attention to. This is really no different from the problem of post hoc analyses, or data-peeking, such as the practice of rummaging through subgroups either via interaction terms (or worse, analyzing subgroups separately) after an experiment is conducted and then interpreting only the largest effects. If you use a sample to construct a model, or to choose a hypothesis to test, you cannot make a rigorous scientific test of the model or the hypothesis using that same sample data. This, by the way, is the real statistical meaning of the term *post hoc*—it does not refer to *afterward* in terms of time. Rather, it refers to looking at the data to decide which tests or parameters will be included in the analysis and interpretation.

To illustrate further the notion of too many unknowns for the sample size, I conducted a small simulation study to show just how easy it is to produce a good-looking model when the data are overfitted. I created 16 noise variables using a random number generator, and arbitrarily chose 15 of those variables to be predictors of the 16th in a multiple regression model. Because all the variables are composed entirely of random numbers, the true model is one in which no real relations exist among any of the variables. Thus, anything we see in the model that looks like a systematic relation between the predictors and response is, by definition, an artifact. In the first simulation, I drew 10,000 random samples for each condition of $N = 50$, $N = 100$, $N = 150$, and $N = 200$. Because there were 15 variables in the models, these sample sizes correspond to ratios of 3.3, 6.6, 10,

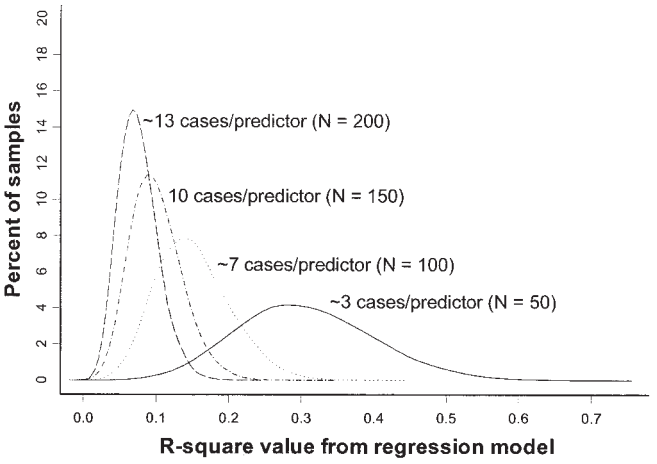

Figure 2. Pure noise variables still produce good R^2 values if the model is overfitted. The distribution of R^2 values from a series of simulated regression models containing only noise variables. The model contained 15 predictors, each consisting of randomly generated values, and a response variable, whose values were also randomly generated. Thus, the true model has an R^2 of 0. Four sets of 10,000 random samples were drawn, each of sample size $N = 50$, $N = 100$, $N = 150$, and $N = 200$. The smoothed frequency distribution of the R^2 values generated by each of the 10,000 models is plotted here for the 4 sample size conditions. Note that even when the number of cases per predictor is reasonably good (200/15=13.3), there are, solely because of the chance of the draw, a fair number of non-0 R^2 values. When there were only approximately 50/15=3.3 observations per predictor, the frequency of large R^2 values was quite high.

[2]I use the term *degrees of freedom* throughout this article in the sense of the number of unknowns in a system that are free to vary. The number of degrees of freedom available for estimation in a regression model is roughly the number of unique bits of information (observations or cases) minus the number of unknowns (regression weights) to be estimated. In other words, each unknown uses up a degree of freedom. The notion of using too many degrees of freedom given the amount of information is a key idea in many branches of science and is directly related to the concepts of parsimony and falsification, or disconfirmation, in the philosophy of science. See Mulaik (26), for example.

OVERFITTING IN REGRESSION-TYPE MODELS

and 13.3 observations per predictor, respectively. I collected the R^2 values for each model and plotted them in Figure 2. The plot shows very clearly that the rate of spurious R^2 values increases considerably as the ratio of observations per predictor becomes smaller. In other words, if you put enough predictors in a model, you are very likely to get something that looks important regardless of whether there is anything important going on in the population.

For many decades, there have been a variety of rules of thumb for the sample sizes required for modeling. As it turns out, recent simulation studies have shown that they are not all that bad. We should always bear in mind, however, that such rules are only approximations, and that situations will arise in which we need fewer, but more likely, more observations than they suggest. For linear models, such as multiple regression, a minimum of 10 to 15 observations per predictor variable will generally allow good estimates. Green (7) showed that in terms of the power to detect a typical effect size seen in the behavioral sciences, a somewhat better rule might be to have a minimum base sample size of 50 observations and then roughly 8 additional observations per predictor. As Green (7) points out, however, a much larger number of observations may be needed if the effect size is small or the predictors are highly correlated. On the other hand, if the effect size is very large, smaller sample sizes may be sufficient. For binary and survival models, the size of the entire sample is not directly relevant. Rather, it is the limiting sample size that matters. In the case of models with a binary response, if the number of events is smaller than the number of nonevents, the limiting sample size is the number of events. Conversely, if the number of nonevents is the smaller of the 2, we use the number of nonevents as the limiting sample size (in other words, the limiting sample size is $N \times \min[p, 1\text{-}p]$, where p is the proportion of events). For survival models, the limiting sample size is simply the number of events. Thus, even if the sample size is 1000, if there are only 10 events, the limiting sample size is only 10.

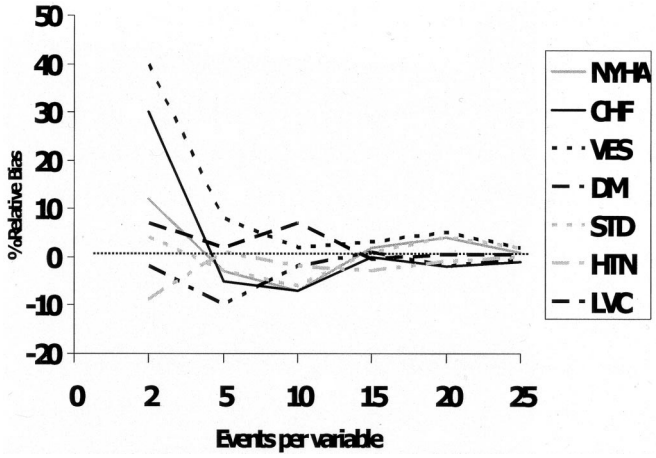

Figure 3. Results of the simulation study of logistic regression models by Peduzzi et al. Pedduzi et al. (9) studied the stability of logistic regression coefficients under a variety of events-per-predictors ratios. Recall that the limiting sample size for a logistic model is the number of events (when there are fewer events than nonevents). The x-axis represents the ratio of events per predictor in the model for the case of 7 predictors. The y-axis shows the percent relative bias in the regression weight compared with the known population weight. The results suggest that bias is unacceptably high when there are fewer than 10 to 15 events per predictor. Reproduced with permission (9).

Peduzzi et al. (8,9) have published simulation studies suggesting that logistic and survival models will produce reasonably stable estimates if the limiting sample size allows a ratio of approximately 10 to 15 observations per predictor. Figure 3 is reproduced from the study by Peduzzi et al. (9) of logistic regression. The simulation used a model developed on real data in which death in cardiac patients was predicted from a variety of medical background variables. Allowing the model derived from this data to represent the true population model, they examined, among other things, how well the true regression weights were recovered depending on the number of events per predictor.

The x-axis in Figure 3 represents the ratio of events per predictor variable in the model, whereas the y-axis shows the percent relative bias of the average regression coefficient from the simulation samples compared with the true coefficient. The results suggest that when there were fewer than 10 events per predictor, the estimates tended to be badly biased. In fact, one might argue based on the figure that the ratio should be at least 15. In further analyses from the same article, Peduzzi et al. (9) show that in many cases, this bias was not trivial in magnitude. For example, when there were only 5 events per predictor, anywhere from 10% to 50% of the simulated regression weights were biased by as much as 100%. When only 2 events per predictor were available, 30% to 70% of the estimates had bias greater than 100%! At a minimum, these results should make us very wary of an article that does not at least meet the rough guideline of 10 to 15 events per predictor—an all too common feature of many published articles. These results should also give us great pause when we plan our own studies and conduct analyses. As noted, Peduzzi et al. (8) have also shown that the rule of 10 to 15 events per predictor applies to survival analyses. As was the case for linear regression, there are conditions under which we might need more events for logistic or survival models, such as small effect sizes, truncated ranges in the predictor variables, or extreme event probabilities. Thus, in designing a study, we should plan a sample size that will allow us to estimate the most complex model we might be interested in (in terms of the number of predictors and also nonlinear, multiplicative terms, or subgroup analyses), such that the estimates will be stable and the fit not overly optimistic. If we cannot gather a sample of sufficient size, we have to find ways to simplify our model, ie, use fewer degrees of freedom, or we have to correct for overfitting. Before we turn to these other approaches, we should focus on several common techniques that actually make matters worse.

M. A. BABYAK

WHAT NOT TO DO TO PRESERVE DEGREES OF FREEDOM

Automated Stepwise Regression[3]

Many modeling software packages include an option that automatically selects the best model. The best-known forms of automated selection are forward or backward stepwise selection. In a nutshell, these are algorithms that sift through large numbers of correlations and partial correlations, fitting and refitting models until some final model is achieved. Researchers typically conduct this type of regression analysis by specifying some keyword such as *backward*, *forward*, or *stepwise*, depending on the software. They allow the algorithm to churn and produce a final model, which they in turn usually report as-is in the final published article. The problems with automated selection conducted in this very typical manner are so numerous that it would be hard to catalogue all of them here. One psychology journal has even made it an editorial policy to summarily reject any article that uses the technique (10). Nevertheless, the practice remains surprisingly popular, and such models continue to appear in even the most prestigious journals, in some cases using the results to draw conclusions about life-and-death clinical issues. Although automated selection is still discussed in some popular statistics textbooks, most advised caution in its use even in the early days of its development, when excitement abounded about the possibility that machines alone would be able to solve our problems. For many, it was always intuitively evident that letting the machine think for you was not a good idea, and certainly we've long understood the problems created by conducting many, many statistical tests (which is what automated selection does). The primary problem with automated selection is that under the most typical conditions we see in medical and psychological research, ie, moderate-to-small sample sizes and many possible predictors, many of which are correlated with one another, the possibility of overfitting is far too great for the results to have anything but the most tentative interpretation. The late Jacob Cohen perhaps put it best when he said that the "capitalization on chance that [stepwise selection] entails is more than I know how to compute" (11). Although it may look like we have not used many degrees of freedom in the final model, we have actually used up a whole passel of them along the way during the selection process. These phantom degrees of freedom just happen to be hidden from us at the end stage.

More recently, simulation studies have supported the earlier warnings about automated selection, showing that, unless special corrections are made, the problem of overfitting can be quite grave in automated regression (12,13). The simulations of Derksen and Keselman (13) are among several that demonstrate just how poorly automated selection performs under conditions that are frequently encountered in real-world research. They found that stepwise selection produced final models in which 30% to 70% of the predictors were actually not related to the response in the population, ie, were pure noise. They further concluded that: a) "The degree of correlation between candidate predictors affected the frequency with which the authentic predictors found their way into the model," b) "The greater the number of candidate predictors, the greater the number of noise variables were included in the model," and c) "Sample size was of little practical importance in determining the number of authentic variables contained in the final model." More recently, Steyerberg et al. (14) simulated a series of logistic regression models that compared automated stepwise selection, forced simultaneous entry (all candidate variables remain in the model), and a "sign-correct" method (removing variables whose regression coefficients did not have a sign that made substantive sense, eg, age inversely related to death). Automated selection proved to be the least desirable procedure among the 3 with respect to overfitting, even when the other 2 procedures exceeded the recommended events per predictor guidelines. It is not only the traditional stepwise procedures that overfit models. Even some of the more sophisticated alternatives such as best subset regression do not solve the problem of overfitting (15), again because more degrees of freedom are used than the sample size can support. I should not leave this topic without adding that procedures for correcting these particular overfitting problems have existed for many years (16)—it has been known for some time, for example, that the correct model degrees of freedom for stepwise procedures is really closer to the total number of all candidate predictors—but these corrections are apparently almost uniformly ignored by most researchers in our field. The bottom line here is that if an article reports the results of a regression model that has used an uncorrected stepwise selection process, be extremely skeptical of the conclusions. The model and consequent conclusions may indeed be correct—but there is simply no way of being certain. Automated selection algorithms that do make the appropriate corrections, such as the lasso method by Tibshirani (17), have been developed, but they are not yet widely available. If you absolutely insist on using a stepwise algorithms, simulations by Steyerberg et al. (14) showed that the least harmful of the approaches is probably backward selection using a very liberal p value criterion, say .50, again paying careful attention to the guidelines for sample size and the number of predictors. The authors argue that this is probably because when we have too many candidate predictors for the sample at hand, there is not much power to detect important predictors. The liberal criterion p value compensates by making it more likely that truly important predictors will be retained in the model. However, there is still the problem of many unimportant variables being included in the final model, but these simulations suggest that, under these particular circumstances, the inclusion of true predictors via the liberal entry criterion outweighs the problem of including unimportant variables.

[3]Sometimes the term *stepwise regression* is used to refer to hierarchical or block forced entry regression, in which variables are forced and maintained in a model sequentially, as described in, for example, Cohen and Cohen (18). This technique is quite distinct, however, from automated stepwise regression, and, when conducted properly, is not associated with the problems I describe.

OVERFITTING IN REGRESSION-TYPE MODELS

Univariate Pretesting or Screening

One very common way of selecting variables for a regression model is to look at the univariate relation between each variable and the response, and then to cull only those variables significant for entry into the subsequent regression analysis. Although it may appear innocent enough, this is actually just a variant of automated selection in disguise; even though we did the testing in a preliminary manual step, we have still spent degrees of freedom against the sample and increased the risk of overfitting. The true degrees of freedom for the regression model should be the total number spent in all stages of data analysis. Using univariate prescreening also creates other problems in the context of multivariable modeling. For example, variables in isolation may behave quite differently with respect to the response variable when they are considered simultaneously with 1 or more other variables. If there is suppression (18), for example, the relation between a variable and outcome may not appear to be important at all in the univariate case, but may become quite important after adjustment for other covariables. Pretesting using *p* values is not the only culprit in producing these so-called phantom degrees of freedom. For example, many of us are taught to explore the possibility of a nonlinear relation between a predictor using a 2-step procedure. In the first step, we include a linear and nonlinear term, eg, x and x^2, in the model. If the nonlinear term is not significant, then we remove it from the equation and report only the linear term result. Grambsch and O'Brien (19), however, show that when we test and then remove a nonsignificant nonlinear term and report on only the linear coefficient, the remaining linear test is too liberal unless we account for the degree of freedom spent in testing that nonlinear term. Extending this thinking beyond traditional pretesting, Faraway (20) demonstrated that these phantom degrees of freedom actually arise in all sorts of unexpected places, such as examining residuals for homogeneity of variance, testing for outliers, or making transformations to improve power, to name a few, underscoring the principle that virtually any data-driven decision about modeling will lead to an overly optimistic model. The article by Faraway (20) in particular has rather depressing implications for how we go about much of our data analysis in many fields, including our own. For example, it shows that there is only so much information that can reliably extracted from a given data set before we have used up far too many phantom degrees of freedom. Statisticians have long recognized something like this uncomfortable complication with respect to multiple tests and the increase in Type I error—often joking that if one took the most conservative stance with respect to multiple testing, we would probably have to be allocated a certain number of tests per career, retiring when we had expended them all. If there is a limit to how many degrees of freedom we can use with a given sample, how should we proceed, especially with respect to the usual practice of the repeated analyses of public data sets? I suppose a reasonable approach might be the same as many recommend with respect to the multiple testing problem: if a researcher enters the analysis in good faith with a firm set of a priori hypotheses or tests in mind, and stays within the guidelines for sample size for a given model, we can at least be assured that the overfitting problem has been at least contained, if not perfectly controlled, for that particular model.

Dichotomizing Continuous Variables

It may appear strange to introduce this topic here, but in some very commonly encountered circumstances, categorizing a continuously measured predictor also will lead to overly optimistic results. Most readers should be familiar with the idea that, apart from its absurdity from a measurement perspective, chopping variables like blood pressure or age into groups will necessarily result in a loss of information, lower measurement precision, and usually a considerable loss of power in subsequent analyses.[4] What is less well known is that the common practice of dichotomizing 2 continuous variables and using them as factors in an ANOVA will yield an unacceptable Type I error rate when those 2 original variables are even moderately correlated. Because ANOVA is just a special case of the general linear model, this problem also will haunt us in the multivariable regression situations. Maxwell and Delaney (21) studied what happens when the continuous predictors, x_1 and x_2, are dichotomized at the median and the true model is $y = 0.5x_1 + 0x_2$. In other words, in the population the continuous variable x_1 is related to y, but the continuous variable x_2 is not. The trouble arises when x_1 and x_2 are correlated, and gets worse as that correlation increases. Table 1 shows that the Type I error rate associated with the relation between x_2 and y increases dramatically as a function of the correlation between x_1 and x_2. The key to interpreting this table is that the true relation between x_2 and y is 0, so that if the model were fitted appropriately, the Type I error rate should be something like 0.05 for the test of the relation between x_2 and y. In the first column of Table 1, we can see that regardless of sample size, when x_1 and x_2 are uncorrelated, an acceptable Type I error rate is preserved, ie, we reject the null hypothesis that x_2 has a non-0 relation with y only approximately 5% of the time over the long run. When the correlation between x_1 and x_2 increases, however, dichotomizing both variables begins to increase the likelihood of a Type I error. When the correlation between x_1 and x_2 exceeds 0.5, the Type I error rate becomes alarmingly high. Maxwell and Delaney (21) further demonstrate that dichotomization also can yield spuriously significant interactions if there is a nonlinear relation between 1 of the predictors and the outcome. It probably is not hard at all to find an article in the published literature in which 2 (or more) correlated variables, such as depression and hostility, body mass index and blood pressure, or perhaps 2 related aspects of job stress, have been dichotomized and entered into a linear model. As in the case with models that have used too many degrees of freedom for the

[4]For a thorough and damning review of dichotomization in general, see MacCallum et al. (27). Readers also may find the simulation applet by McClelland (28) very useful in developing an intuitive understanding of how dichotomization reduces power.

TABLE 1. Type I Error Rates for the Relation Between x_2 and y After Dichotomizing 2 Continuous Predictors[a]

N	Correlation Between x_1 and x_2			
	0	.3	.5	.7
50	.05	.06	.08	.10
100	.05	.08	.12	.18
200	.05	.10	.19	.31

[a] Maxwell and Delaney (21) calculated the effect of dichotomizing 2 continuous predictors as a function of the correlation between them. The true model is $y = .5x_1 + 0x_2$ where all variables are continuous. If x_1 and x_2 are dichotomized, the error rate for the relation between x_2 and y increases as the correlation between x_1 and x_2 increases. This table is reproduced with permission.

sample size, dichotomization is yet another way to produce results that do not really reflect the true model.

Multiple Testing of Confounders

A prime assumption of any regression-type model is that we have included all of the important variables in the equation. Because of constraints in resources and the limitations of our knowledge, this assumption is, of course, virtually impossible to meet in most applied research settings, particularly with observational designs. What we can do, at a minimum, is try to collect data on suspected confounders and ensure that we have made some account of their effect by including them in the model. If we can afford it, we also can improve the precision of the model by including variables that are not necessarily related strongly to the other predictors (and hence not confounders), but that are related to the response. Finally, if we are interested in potential mechanisms that might help explain the relation between some predictor and the response, we can include those putative mediating variables in the model and determine whether the results are consistent with our causal hypotheses. Being able to include extra explanatory variables or possible mediators in a model is nice, but the concern about confounders is one that we recognize as most immediate. Indeed, having published a fair number of regression-type model results, I can tell you that these publications, regardless of whether they are based on randomized trials or observational data, are almost inevitably followed by a flow of personal communications and published correspondence pointing out which covariates I have failed to include in the model. In most cases, if I had included all of the suggested covariates and also wanted to follow the sample size-to-predictors guidelines discussed earlier, I would spend the rest of my career waiting for enough data to be collected. I am exaggerating (though in some cases only slightly), but it does highlight the point that you have to understand the limitations of your data and often make some hard choices, including accepting the possibility that there always may be a lurking confounder out there somewhere, vs. living with an overfitted model with biased coefficients.

If you do have a pool of potential confounders measured and available for analysis, but do not have the sample size to support that many tests, what is the best way to proceed? This is a very difficult and controversial question, and inevitably spurs lively debates among statisticians. The conventional practice is, irrespective of sample size, to add 1 or more of the potential confounding variables to the model, and if the effect of the predictor of interest is wiped out, we conclude that the original relation was confounded with the variables we just added. However, in the context of the phantom degrees of freedom problem, we should immediately see that there are pitfalls to this practice. First, for each new predictor variable we add to the model, we again have expended a degree of freedom, regardless of whether that new variable ultimately appears in the final model or not. As discussed, the number of degrees of freedom expended, whether explicit or phantom, should be kept within the limits of the sample size, or, as we have seen earlier, we will not be able to trust the final model. Second, because the correlations among the predictor, the putative confounder, and the response are all subject to sampling error, one of the confounders will knock the predictor out of the model by chance alone if you test enough of them. The real problem here is that unless you have been very careful to account for expended degrees of freedom, you will not have any way of knowing the extent to which the apparent confounder is a real confounder or just caused by the play of chance sampling. As I noted earlier, the dilemma boils down to whether we are more concerned about confounders or about deriving an overfitted model. My personal preference is to choose a priori a set of predictors whose number or complexity remains within the sample size limits discussed and to stay with that model no matter what. This does risk overlooking a hidden confounder that I have not anticipated, but on the other hand, the resultant model is entirely transparent to me and the scientific community: there are no phantom degrees of freedom or overfitting, and consequently, the model will be more likely to replicate in new samples. What if I am still concerned about confounders? I have always been of the mind that the typical research report should be divided into 2 sections, 1 for a priori hypotheses and another for "interesting stuff I've found in this data set that may or may not be reliable—I just can't be sure." I would put the confounder search in the latter section and use very tentative language to interpret the results. Alternatively, one can adopt some of the strategies suggested to preserve degrees of freedom in the section below, or apply the appropriate correction for the number of degrees of freedom really expended in the analysis (see the section on shrinkage and penalization likelihood below), and then have the luxury of interpreting the results with more confidence.

I am aware that this is a controversial position. In fact, after reading a first draft of this manuscript, 1 reader understandably posed the semiserious question, "Why should I be punished for looking for confounders?" I teased back that I like to think of it more of a sobering-up than a punishment! At a minimum, I hope that researchers will give some thoughtful consideration to the issue of overfitting because of multiple

confounder searches, perhaps limiting the number of tests they perform, applying the recommended techniques or corrections discussed below, or at least couching their interpretation of the results in appropriately cautious language. If nothing else, this dilemma underscores the need to plan a study such that there are sufficient observations to include in the model based on the number of confounders they think might be important to include. Perhaps someday there will be a series of simulation studies that will help illuminate this specific problem further. At present, I know of no study that addresses this question directly.

WHAT TO DO INSTEAD: SOME STRATEGIES FOR AVOIDING OVERFITTING

Collect More Data

Sometimes we have to come to grips with the reality that we simply do not have enough data to answer a given question. We just have to get out and collect more data, and all the sophisticated technical fixes in the world will not change that fact. For example, if there are only 20 events available for a survival analysis, and we absolutely need to estimate the regression coefficients for even just 5 or 6 predictors, we know that these are not really enough data to provide much in the way of stable estimates. We are therefore better off finding a way to gather more events, either by increasing the follow-up time or by recruiting more participants into the study. If this is not possible, there might be some heuristic value in reporting something from the available data, but it should be made extremely clear that the result is tentative at best.

Combine Predictors

One obvious approach to preserve degrees of freedom is to reduce the number of predictors in the model. As we have learned, however, removing a predictor by peeking at its relation with the response, either through a mechanized procedure or by hand, generates problems. Ideally, we would select a limited number of predictor variables based on our substantive knowledge and maintain them in the model regardless of how it turns out. Sometimes, however, we do not have enough information to make reasonable a priori choices about predictors and are left with using the data at hand to choose variables to include in the model. I now briefly discuss a few possible approaches for reducing the number of variables in the model in ways that avoid or minimize overfitting. Notice that all of these approaches avoid peeking at the data with respect to the relations or tests of interest—that is, they do not rely on examining the relation between the predictors and the response variable to select which variables to include or exclude from a model.

One useful, easy approach is to combine or eliminate closely correlated predictors. For example, it might be possible to use a clustering algorithm, such as recursive partitioning, factor analysis, or principal components, to combine 2 or more predictors into 1 variable (of course, in a hypothesis-testing context, we would combine only variables that are of secondary importance to our main question). Alternatively, some covariables might be combined into a single score based on theoretical knowledge or previous results. For example, for a model in which all-cause mortality is the response, we might use 1 of the recognized comorbidity indexes (22) for weighting several medical and demographic covariates to combine those variables into a single composite, thus saving many degrees of freedom. When the weights for these indices are derived from very large samples, the resulting composite is probably a better estimate of the true relation between the component variables and the outcome in question in the population than the estimate we might make from our smaller sample. Using an index still results in a tradeoff—we necessarily lose specific information about the components of the index but are preserving degrees of freedom in the model. The decision to use an index at all, of course, also rests on how confident we are in its theoretical and measurement validity.

As an alternative to using indexes as a means of preserving degrees of freedom, we might instead investigate the possibility of fixing some regression coefficients if the relations between that predictor and the response in question have been well-studied. For example, if we see again and again in the literature that age produces a consistent risk ratio with respect to the probability of a cardiac event, we might be confident enough about this relation to make the regression coefficient for age a constant in the model, thus saving a degree of freedom. This approach is taken in many fields, but the practice has not made its way into psychosomatic research. Econometricians, for example, often fix some, or even all, of the regression coefficients in a model because they are willing to assume that the values derived over years of study are probably about right, or at least good enough to serve as an adjustment variable in the model. In other words, they are willing to make those values an assumption of the model. If our study sample is similar enough to those from which these indexes or single fixed coefficients have been computed, we might be willing to trade the loss of sample-specific information about these covariates for the reduction in bias associated with a better sample size-to-predictors ratio.

Shrinkage and Penalization

Even after you have accounted for phantom degrees of freedom and avoided asking more from the data than the sample size can support, the final model can still be too optimistic. (And as we have seen, if we haven't paid attention to these issues, we know that the model will be too optimistic). Shrinkage techniques allow us to understand the extent of this overoptimism and generate an estimate of how well the model might fit in a new sample. The adjusted R^2 that appears on the output of many statistical packages is actually a type of shrinkage estimator. This value is an estimate what the fit of the regression model would be if it were fitted against a new data set (assuming that you already have accurately accounted for all the degrees of freedom.) Computing power also has led to newer approaches to shrinkage. For example, a technique called *bootstrapping* (23) can generate estimates of shrinkage

not just for the fit of the model but also for many other aspects of the model, such the regression weights and intercept. Bootstrapping is actually a variant of simulation, with the important distinction that repeated samples are drawn with replacement from the data set at hand. As in simulation, the model under study is then estimated for each sample and the results are tabulated. We can, for example, derive a distribution of R^2 values, b-weights, means, standard deviations, and so forth from the repeated samples. (Many statisticians I have spoken with believe that resampling techniques such as bootstrapping will probably replace traditional theory-based test statistics some day, such that we will be reporting empirical p values rather than traditional ones derived from the theoretical sampling distributions.) Bootstrap validation of models also has been shown to be superior to older techniques of model validation, such as splitting the data set into training and testing halves (24). Even when bootstrapping is not included in the software package (S-Plus and R, for example, include modules), it is relatively easy to implement, and the code for many popular software packages such as SAS and SPSS is readily available on the Internet. I imagine that bootstrapping will become a standard part of many statistical software packages in the next few years. Advanced computing also has allowed the relatively easy calculation of even more complex shrinkage algorithms, such as maximum likelihood penalization (see Harrell [3], pp. 207–210). This latter approach is a sort of preshrinking of the regression coefficients and fit, such that the resultant model will be much more likely to replicate. Penalization has the advantage of allowing us to adjust specific areas of the model in which complexity (eg, interaction terms, nonlinear terms) may have given us a fit that we really did not deserve. Steyerberg et al. (14) have shown that models using preshrunken estimates and a fixed set of predictors tend to be the most likely to replicate in new samples. Unlike bootstrapping, with the exception of S-Plus, penalized likelihood estimation is not yet widely available in canned form in the major statistical software packages.

CONCLUSION

In the preceding pages, I have covered a relatively narrow but important aspect of model fitting. The broad essence of what I have tried to convey is that we need to be mindful of the known limitations of our analytic tools, and that there is still no substitute for thinking long and hard about the scientific question at hand (see Freedman [25] for an excellent discussion of this concept). Good analytic practice, of course, requires a whole host of considerations not discussed here, such as careful attention to the reliability and validity of the measures, the selection of the sample, the range of the predictors, and satisfaction of a number of other underlying assumptions, to name a few. A quick perusal of our literature suggests, however, that overfitting is not a bad place to start in terms of putting our scientific house in better order. As a field, we have been overfitting models for years, in some cases, very badly. In some cases, this may be of little consequence. However, it is not hard to imagine that millions of research dollars and uncountable hours of work are spent each year chasing findings or ideas that arose from the failure to appreciate this concept more fully. Perhaps worse, given the prevalence of overfitted models, some of these spurious conclusions must surely have made their way into the world of clinical decision-making. There also may be ramifications for the credibility of the science among the public. One study shows that variable x is a risk factor for heart disease (and, of course, the press makes much of it), whereas the next study repudiates the variable x theory, showing instead that variable x is not a risk factor at all—it is really variable z! Which do we believe? Although there are a number of reasons that findings tend to fluctuate across studies, tightening up our modeling practices might go a long way in reducing the frequency with which this confusion arises.

As I noted at the outset of this article, none of this is meant to suggest that we should be slavishly conservative in how we test our scientific ideas. On the contrary, we have an obligation to entertain and explore every sort of means of understanding of our precious data. However, we have an equal duty to understand and appreciate the varying shades of rigor associated with each of these pursuits and to interpret and report results accordingly. It is my hope that this article has inspired readers to consider some of these issues more deeply, perhaps digging into the literature on their own or with colleagues, perhaps initiating debate, or even incorporating some of the points outlined into their own research endeavors.

I am extremely grateful to Frank Harrell, PhD, Beverly Brummett, PhD, and Heather Lett, MA, for their very helpful comments on earlier drafts of this manuscript.

REFERENCES

1. Taleb NN. Fooled by randomness. New York: Texere; 2001.
2. Mendes de Leon CF, Babyak MA. Advanced quantitative methods in psychosomatic research. Workshop presented to the Annual Meeting of the American Psychosomatic Society, Phoenix, AZ, 2003.
3. Harrell FE Jr. Regression modeling strategies: with applications to linear models, logistic regression and survival analysis. New York: Springer; 2001.
4. McCullagh P, Nelder J. Generalized linear models. London: Chapman & Hall; 1989.
5. Cox DR. Regression models and life tables. J R Stat Soc B 1972:187–202.
6. Hardin JW, Hilbe JM. Generalized estimating equations. London: Chapman & Hall/CRC; 2003; 34.
7. Green SB. How many subjects does it take to do a regression analysis? Multivar Behav Res 1991;26:499–510.
8. Peduzzi PN, Concato J, Holford TR, Feinstein AR. The importance of events per independent variable in multivariable analysis, II: accuracy and precision of regression estimates. J Clin Epidemiol 1995;48:1503–10.
9. Peduzzi PN, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. J Clin Epidemiol 1996;49:1373–9.
10. Thompson B. Stepwise regression and stepwise discriminant analysis need not apply here: a guidelines editorial. Ed Psychol Meas 1995;55:525–34.
11. Cohen J. Things I have learned (so far). Am Psychol 1990;45:1304–12.
12. Altman DG, Andersen PK. Bootstrap investigation of the stability of a Cox regression model. Stat Med 2003;8:771–83.
13. Derksen S, Keselman HJ. Backward, forward and stepwise automated subset selection algorithms: frequency of obtaining authentic and noise variables. Br J Math Stat Psychol 1992;45:265–82.
14. Steyerberg EW, Harrell FE, Habbema JD. Prognostic modeling with logistic regression analysis: in search of a sensible strategy in small data sets. Med Decis Making 2001;21:45–56.

OVERFITTING IN REGRESSION-TYPE MODELS

15. Roecker EB. Prediction error and its estimation for subset-selected models. Technometrics 1991;33:459–68.
16. Copas JB. Regression, prediction, and shrinkage (with discussion). J R Stat Soc B 1983;45:311–54.
17. Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc B 2003;58:267–88.
18. Cohen J, Cohen P. Applied multiple regression/correlation analysis for the behavioral sciences. Hillsdale, NJ: Lawrence Erlbaum Associates; 1983.
19. Grambsch PM, O'Brien PC. The effects of preliminary tests for nonlinearity in regression. Stat Med 1991;10:697–709.
20. Faraway JJ. The cost of data analysis. J Comput Graph Stat 1992;1:213–29.
21. Maxwell SE, Delaney HD. Bivariate median splits and spurious statistical significance. Psychol Bull 1993;113:181–90.
22. D'Hoore W, Sicotte C, Tilquin C. Risk adjustment in out-come assessment: the Charlson Comorbidity Index. Methods Inf Med 1993;32:382–7.
23. Efron B, Tibshirani R. An introduction to the bootstrap. London: Chapman & Hall; 2003.
24. Steyerberg EW, Harrell FE Jr, Borsboom GJ, Eijkemans MJ, Vergouwe Y, Habbema JD. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. J Clin Epidemiol 2001; 54:774–81.
25. Freedman D. Statistical models and shoe leather (with discussion). Soc Methodol 1991;21:291–313.
26. Mulaik SA. The metaphoric origins of objectivity, subjectivity and consciousness in the direct perception of reality. Philos Sci 1995;62:283–303.
27. MacCallum RC, Zhang S, Preacher KJ, Rucker DD. On the practice of dichotomization of quantitative variables. Psychol Methods 2002;7:19–40.
28. McClelland G. Negative consequences of dichotomizing continuous predictor variables. Available at: http://psych.colorado.edu/~mcclella/MedianSplit/.