Independent Predictors from Stepwise Logistic Regression May Be Nothing More than Publishable *P* Values

Nathan L. Pace, MD, MStat

Observational data consisting of patient records extracted from large databases have become a fertile source of material for epidemiological reports on diseases, treatments, events, and outcomes.[1] Analysis of such data can assess the association between exposure and outcome, can construct models for prediction of prognosis, and can generate hypotheses for further clinical research. Some of the data sets can be very extensive, such as the more than 800,000 records of the National Surgical Quality Improvement Program of the Department of Veterans Affairs that were used by Bishop et al.[2] to relate day-of-surgery deaths to the available covariates. These covariates included binary variables such as sex or history of peripheral vascular disease, ordinal variables such as ASA physical status, nominal variables such as type of surgery, and continuous variables such as serum creatinine concentration. Bishop et al. entered 45 covariates into a logistic regression analysis to identify "independent predictors" of mortality. The analysis identified 17 statistically significant covariates (predictors) of day-of-surgery mortality. The model is purely descriptive of the observed data and is not a mechanistic or pathophysiologic explanation of the effect of the covariates on outcome. The purpose of this editorial is to highlight some difficulties and controversies about logistic regression of observational data with particular focus on the process of automatic variable selection with stepwise logistic regression.

Multivariable linear regression is the creation of a linear model relating some continuous response variable y (e.g., blood pressure) to a collection of k explanatory variables, also called covariates. This regression starts with n independent observations (n patients) of the form (y_i, $x_{i,1}$, $x_{i,2}$,…, $x_{i,k}$) where the subscript i denotes the ith patient, y_i is the observed value in the ith patient, x_i is the value of a covariate in the ith patient, and the second subscript of x_i denotes the indices of observations of the k covariates. The linear model equation for the ith individual is

$$\mu_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_k x_{i,k} = \sum_{j=0}^{k} \beta_j x_{i,j}$$

with j being the index of the kth covariate. The β_js are the unknown coefficients (parameters) of the model that will be estimated from the observed data. For the ith individual, the expected value of the model, μ_i, is the linear sum of each covariate value multiplied by its coefficient. The difference between the observed (y_i) and expected value (μ_i) reflects biological variability, measurement error, etc. This model can be made more complex if interactions such as $x_j x_j$ or x_j^2 are included. An example of an interaction would be synergy between increasing age and heavier smoking to increase blood pressure. A model including interactions of covariates or quadratic powers of covariates is still a linear model, being linear in parameters. Under least squares regression, the estimation of the parameters of a multivariable linear model generally has an exact closed form solution.

If the response variable is binary, multivariable regression has been extended to create generalized linear models. The most commonly used

This article has supplementary material on the Web site: www.anesthesia-analgesia.org.

From the Department of Anesthesiology, University of Utah School of Medicine, Salt Lake City, Utah.

Accepted for publication August 14, 2008.

Supported solely by university salary.

Address correspondence and reprint requests to Nathan L. Pace, MD, MStat, Department of Anesthesiology, University of Utah, 30 North 1900 East, 3C444, Salt Lake City, UT 84132-2304. Address e-mail to n.l.pace@utah.edu.

DOI: 10.1213/ane.0b013e31818c1297

format is a multivariable logistic regression model.z This regression also starts with n independent observations of the form $(y_i, x_{i,1}, x_{i,2}, \ldots, x_{i,k})$. In this model, y_i denotes the value of the dichotomous (binary) outcome in the ith individual and is generally coded as 0 or 1 representing the absence or presence of an event (e.g., day-of-surgery mortality). Letting π represent the probability that the response variable has value 1, then the logit transformation, log of the odds ratio $= \ln\left(\frac{\pi}{1 - \pi}\right)$, allows π to be expressed as the linear combination of the covariates without being inconsistent with the laws of probability. The logit transformation is the link function of the logistic generalized linear model; the model equation for the ith individual is

$$\ln\left(\frac{\pi_i}{1 - \pi_i}\right) = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_k x_{i,k} = \sum_{j=0}^{k} \beta_j x_{i,j}.$$

The inverse of the logit transformation,

$$\pi_i = \frac{e^{\beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_k x_{i,k}}}{1 + e^{\beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_k x_{i,k}}} = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_k x_{i,k})}},$$

calculates the predicted probability for each set of covariate values. As each y_i can only assume the values 0 and 1, there will always be a difference between each observed (y_i) and expected value (π_i). The estimation of parameters in a logistic regression model does not have an exact solution, but is derived iteratively by the methods of maximum likelihood.

In an observational data set with many covariates, one possible goal of multiple variable regression, either linear or logistic, is to identify a fitted model that provides reasonably precise estimates of the mean response using a parsimonious set of explanatory variables (fewer covariates). In a data set with k covariates, there are $2^k - 1$ candidate models depending on which covariates are included. For example, in a data set with 50 explanatory variables, the fitted parsimonious model might include the 5th, 7th, 8th, 24th, and 50th covariate. In the report by Bishop et al., there were about 50 covariates or 10^{15} or so possible candidate models, which does not include models with interactions of the covariates. This combinatorial explosion makes impossible the estimation of all possible models. If all models could be estimated, information-theoretic statistics offer methods to choose the best models.[3] Under some circumstances there may be substantial knowledge about the phenomenon being observed that offers guidance in selecting a limited number of covariates. This use of prior knowledge is very much the exception rather than the rule in statistical modeling of observational data.

About 50 years ago, an algorithm was proposed for an automatic procedure to select a statistical model where there are a large number of potential explanatory variables and no underlying theory on which to base the model selection. The algorithm has been incorporated into regression analysis for both linear and generalized linear models and is implemented variously in most statistical software packages. There are many variations of automatic variable selection. The forward variant starts with no covariates in the model. At each step, the covariates are considered sequentially. The criteria for inclusion of a covariate are a suitably significant analysis of variance statistic, multiple coefficient of determination, or index of the maximized log-likelihood, such as the Akaike Information Criterion. The covariate that best fulfills the criteria is entered into the model at that step. The backward variant starts with all covariates in the model. Analogously, a covariate is considered for removal from the model at each step. Most commonly, an automatic variable selection process adds (forward selection) and/or removes (backward elimination) covariates from the model at each step; thus, the term "stepwise regression." The algorithm includes a stopping rule when no further covariates are either removed or entered into the model. The application of stepwise regression in the anesthesia literature is almost always for logistic models and follows one of two paths. In the first, the dimensionality of the covariate set is reduced by univariable screening. Each covariate is tested for significance by a simple t-test (continuous variables) or χ^2 test (categorical variables). Covariates having a nominal P value <0.1 to 0.25 are included in the stepwise regression, and covariates with higher P values are discarded. A second, much less common, approach includes all available covariates in the stepwise regression; this approach was used by Bishop et al.

When inspecting the results of stepwise logistic regression, the question of validity arises. Are the identified independent predictors really associated with the outcome? Are the magnitudes of independent predictors estimated without bias? Because of the statistical complications and complexity of stepwise logistic regression, its properties are usually studied by simulation. To illustrate issues concerning validity, two simulations were performed (Appendix).

In the first, mock data sets with 1000 rows of data were created. Observational data sets in anesthesia reports typically have 100 to 1000 cases. Each row of data had 50 random covariates, half binary and half continuous, and a random binary outcome variable. Each covariate and the outcome variable was produced by a random number generator that produced a 1000 element independently and identically distributed number set, either standard normal distribution or binomial distribution centered at 0.5. A univariable test separated the covariates into those significant and not significant at $P < 0.1$. The regression of the binary outcome on the covariates included only those covariates passing the univariable screening. This entire

simulation was itself repeated 1000 times. In this simulation, the use of randomly generated covariates and outcomes ensured that there was no true association between any covariate and the binary outcome.

In the 1000 simulations there were 825 instances with at least one significant covariate at a $P < 0.05$. The number of significant covariates was 1 in 177 instances, 2 in 237 instances, 3 in 190 instances, and 4 in 130 instances; there were 5–8 significant covariates in 91 instances. It is expected that some spurious correlations will arrive by chance. These are the incorrect conclusions (false-positive results). Without the use of simulation, one might not have thought so many were possible! Essentially, the stepwise logistic regression falsely identified in more than 80% of the simulations an appearance of an association between a random binary outcome and one or more randomly created explanatory variables. From pure noise, we have found independent predictors with publishable P values <0.05.

In the second simulation, a larger data set of random numbers was created (100,000 records) with a rare event outcome (500 events) regressed against 50 covariates (half binary and half continuous). This was similar to the observational data used by Bishop et al. that had about 650 deaths in more than 800,000 patient records. This simulation identified eight randomly generated covariates as independent predictors, three having P values <0.05.

These simulation results are consistent with the limitations of stepwise regression, both linear and logistic, that have been reported in the statistics and epidemiology literature for two decades on both simulated and real data sets. For example, in a stepwise linear model, the percentage of chosen covariates that is noise is about 33%–89%[4] or 20%–74%[5] depending on the particulars of the simulation study. Austin and Tu[6] used bootstrap methods to explore the stability of stepwise logistic regression chosen independent predictors of 30-day mortality (about 11%) from a data set of about 5000 patients who suffered an acute myocardial infarction; about 30 covariates with univariable significance were offered for possible inclusion. The models created were unstable, in that about half of the models had 12 or 13 independent predictors, but another half had 8 to 11 or 14 to 19 independent predictors. Additionally, single covariates had widely varying appearance rates in the models, from 10% (previous acute myocardial infarction) to 100% (age). In another simulation study, the use of univariable screening before stepwise logistic regression actually reduced the ability to detect significant covariates.[7] The authors suggested avoiding this step entirely.

The results described here are simply part of a larger problem, namely, the use of the same data set for structural identification (which covariates?), inference (parameter magnitude?), and validation (does the model work?). Statistical theory has much to say about estimating model parameters, but on the bigger problem of uncertainty about the structure of the model it is mostly silent.[8] Stepwise logistic regression produces publishable P values from the estimation step. However, these P values are not sufficient for both estimation and validation of the model. The consequences for stepwise logistic regression include over optimistic (too large) regression coefficients, spuriously narrow confidence intervals on coefficients and predictions, and the inclusion of noise (unrelated covariates) in the model. These shortcomings are so severe that in some biological fields there are calls to abandon the use of stepwise multiple regression.[9]

I am not advocating the dismissal of stepwise logistic regression methods from anesthesia research, nor am I suggesting that the independent predictors identified by Bishop et al. are necessarily incorrect. But this editorial does advocate adhering to statistical rigor in interpreting the results of such modeling. For example, the data set should have at least 10 times as many events as covariates,[10] a criterion fulfilled by the data used by Bishop et al. Adding other techniques to stepwise logistic regression, such as data splitting, cross validation, or bootstrapping, can provide internal validation of a prognostic model.[11] Bishop et al. chose not to do so. There are better indices of regression calibration and discrimination (Cox binary regression) to compare models.[12] Useful guides for the reporting of stepwise logistic regression have been suggested.[13] Tutorials for developing models have been published.[14,15] Some models need external validation of predictions against a new data set with possibly a recalibration of the regression coefficients.[16]

We can learn a great deal from modeling observational data. However, the use of rigorous statistical approaches in reporting and validating the models will help to exclude the disturbing possibility that stepwise linear regression is doing nothing more than generating publishable P values from pure noise.

APPENDIX

Simulations were run in version 2.7.1 of R:A Language and Environment for Statistical Computing (R Foundation for Statistical Computing, Vienna, Austria) which is available as Free Software under the terms of the Free Software Foundation's GNU General Public License. Additionally, functions in the packages MASS (7.2-42) and epicalc (2.7.1.0) were used. (Simulations available at www.anesthesia-analgesia.org)

REFERENCES

1. Fleisher LA, Barash PG. Governmental databases, hospital information systems, and clinical outcomes: big brother or big help? Anesth Analg 1999;89:811–3
2. Bishop MJ, Souders JE, Peterson CM, Henderson WG, Domino KB. Factors associated with unanticipated day of surgery deaths in Department of Veterans Affairs Hospitals. Anesth Analg 2008;107:1924–35
3. Burnham KP, Anderson DR. Model selection and multimodel inference: a practical information-theoretic approach. 2nd ed. New York: Springer-Verlag, 2002

4. Flack VF, Chang PC. Frequency of selecting noise variables in subset regression analysis: a simulation study. Am Stat 1987;41:84–6
5. 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
6. Austin PC, Tu JV. Automated variable selection methods for logistic regression produced unstable models for predicting acute myocardial infarction mortality. J Clin Epidemiol 2004;57:1138–46
7. Sun G-W, Shook TL, Kay GL. Inappropriate use of bivariable analysis to screen risk factors for use in multivariable analysis. J Clin Epidemiol 1996;49:907–16
8. Chatfield C. Model uncertainty, data mining, and statistical inference. J R Stat Soc Ser A Stat Soc 1995;158:419–66
9. Whittingham MJ, Stephens PA, Bradbury RB, Freckleton RP. Why do we still use stepwise modelling in ecology and behaviour? J Anim Ecol 2006;75:1182–9
10. Peduzzi P, 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
11. Altman DG, Royston P. What do we mean by validating a prognostic model? Stat Med 2000;19:453–73
12. Miller ME, Hui SL, Tierney WM. Validation techniques for logistic regression models. Stat Med 1991;10:1213–26
13. Ottenbacher KJ, Ottenbacher HR, Tooth L, Ostir GV. A review of two journals found that articles using multivariable logistic regression frequently did not report commonly recommended assumptions. J Clin Epidemiol 2004;57:1147–52
14. Harrell FE Jr. Regression modeling strategies with applications to linear models, logistic regression and survival analysis. New York: Spring-Verlag, Inc, 2001
15. Harrell FE Jr, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med 1996;15:361–87
16. Steyerberg EW, Borsboom GJJM, van Houwelingen HC, Eijkemans MJC, Habbema JDF. Validation and updating of predictive logistic regression models: a study on sample size and shrinkage. Stat Med 2004;23:2567–86