Methodological Notes

Cerebrovascular Diseases

Cerebrovasc Dis 2013;35:187–193
DOI: 10.1159/000345491

Received: August 7, 2012
Accepted: October 24, 2012
Published online: February 21, 2013

Multivariable Analysis in Cerebrovascular Research: Practical Notes for the Clinician

Gianpaolo Reboldi[a] Fabio Angeli[b] Paolo Verdecchia[c]

[a]Department of Internal Medicine, University of Perugia, and [b]Department of Cardiology, Hospital 'Media Valle del Tevere', AUSL 2, Perugia, and [c]Department of Internal Medicine, Hospital of Assisi, Assisi, Italy

Key Words
Multivariable analysis · Linear regression · Logistic regression · Survival analysis

Abstract
The term 'multivariate analysis' is often used when one is referring to a multivariable analysis. 'Multivariate', however, implies a statistical analysis with multiple outcomes. In contrast, multivariable analysis is a statistical tool for determining the relative contributions of various factors to a single event or outcome. The purpose of this article is to focus on analyses where multiple predictors are considered. Such an analysis is in contrast to a univariable (or 'simple') analysis, where single predictor variables are considered. We review the basics of multivariable analyses, what assumptions underline them and how they should be interpreted and evaluated.

Introduction

The terms 'multivariate analysis' and 'multivariable analysis' are often used interchangeably in medical and health sciences research. However, multivariate analysis refers to the analysis of multiple outcomes whereas multivariable analysis deals with only one outcome each time [1].

As is obvious from the title, we focus on multivariable, not multivariate, analysis. Multivariable analysis is a statistical tool for determining the relative contributions of different causes to a single event or outcome. For example, some factors are associated with the development of cerebrovascular disease, including family history of stroke, older age, high blood pressure (BP), diabetes, overweight, elevated cholesterol levels, interventions and cigarette smoking [2]. Multivariable analysis allows us to determine the independent contribution of each of these risk factors (explanatory variables) to the development of the disease (response variable). In other words, the risk of an outcome may be modified by other risk variables or by their interactions, and these effects can be assessed by multivariable analysis.

In this paper, we introduce the clinician to the different multivariable analyses arising from using different measurement scales for the outcome and we detail the assumptions behind each type of multivariable analysis. We also discuss how the different multivariable analyses are interpreted adding some illustrative examples.

Role of Confounders

The relationship between an event (or outcome measure) and a risk factor may be confounded by other variables. Confounding occurs when the apparent association between a risk factor and an outcome is affected by

KARGER

Fax +41 61 306 12 34
E-Mail karger@karger.ch
www.karger.com

1015–9770/13/0352–0187$38.00/0

Accessible online at:
www.karger.com/ced

Gianpaolo Reboldi, MD, PhD, MSc
Department of Internal Medicine
University of Perugia
IT–06126 Perugia (Italy)
E-Mail paolo@unipg.it

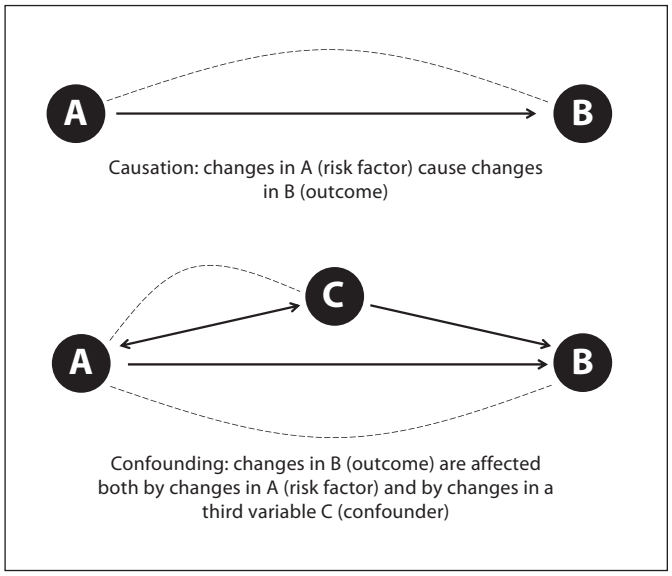

Fig. 1. Relationship among risk factor and outcome (upper panel). The ability of multivariable analysis to simultaneously assess the independent contribution of a number of risk factors to outcome is important when 'confounding' exists (lower panel). Confounding occurs when the apparent association between a risk factor and an outcome is affected by the relationship of a third variable (confounder) to the risk factor and the outcome.

the relationship of a third variable to the risk factor and to the outcome (fig. 1) [1].

A variable cannot be a confounder if it is a step in the causal chain or pathway. For example, moderate alcohol consumption increases serum high-density lipoprotein levels which, in turn, decrease the risk of stroke. In this context, the high-density lipoprotein level is a step in this casual chain, not a confounder that needs to be controlled [1].

A simple and practical technique to assess and eliminating confounding is stratified analysis. Stratified analysis gauges the effect of a risk factor on outcome while holding another variable constant.

As example, consider the relationship between hypertensive target organ damage and risk of stroke [3]. Data from the Progetto Ipertensione Umbria Monitoraggio Ambulatoriale (PIUMA) study [4] allows us to illustrate how to perform a stratified analysis. In particular, we test the association of left ventricular (LV) hypertrophy at ECG with the risk of cerebrovascular events. The crude rates of stroke between hypertensive subjects without or with LV hypertrophy were 0.56 events × 100 patient-years in the former and 1.46 events × 100 patient-years in the latter group. At the univariable survival analysis, the presence of LV hypertrophy at ECG confers an increased risk of future cerebrovascular events (HR 2.37; 95% CI 1.69–3.32; $p < 0.0001$).

However, systolic BP is a potential confounder because it is associated with both LV hypertrophy and cerebrovascular disease [5]. So, we compare the prognostic impact of LV hypertrophy separately among hypertensive patients with systolic BP below and above the median (152 mm Hg). The crude rates of stroke and the risk of cerebrovascular events are greater among hypertensive patients with LV hypertrophy compared to those without LV hypertrophy, both among subjects with systolic BP below and above 152 mm Hg. Stratum-specific HRs (2.84 for patients with systolic BP ≤152 mm Hg and 1.86 for patients with systolic BP >152 mm Hg) largely differ from the HR computed for the entire population (2.34), indicating that there is confounding by systolic BP.

Consider that other variables might affect the relationship between LV hypertrophy and the risk of stroke. In particular, different stratified analyses may prove that the effect of LV hypertrophy on stroke is confounded not only by systolic BP, but also by age, sex and diabetes [4].

To stratify by two variables (for example systolic BP and sex), we need to assess the relationship between LV hypertrophy and stroke in four groups (males with systolic BP ≤152 mm Hg, males with systolic BP >152 mm Hg, women with systolic BP ≤152 mm Hg and women with systolic BP >152 mm Hg). Adding diabetes as variable to the previous stratified analysis, we have eight groups.

For each stratification variable we add, we increase the number of subgroups for which we have to individually assess whether the relationship between LV hypertrophy and stroke holds and we may have an insufficient sample size in some of these subgroups, even if we started with a large sample size. Multivariable analysis overcomes these limitations and it allows us to simultaneously assess the impact of multiple independent variables on outcome.

Common Types of Multivariable Analysis

Multivariable analyses are widely used in observational studies of etiology, intervention studies (randomized and nonrandomized), studies of diagnosis and studies of prognosis [6]. The most common types of multivariable analysis used in clinical research include linear regression, logistic regression and proportional hazard regression (Cox).

Table 1. Multivariable linear regression analyses to test the independent relationship between pulse pressure and other clinical variables

Model/variables	Beta coefficient	SE of the estimate	95% CI	p value
Model 1 (R^2 = 0.56)				
Age (years)	0.594	0.043	0.511–0.678	<0.0001
Glucose (mg/dl)	0.087	0.014	0.060–0.115	<0.0001
ECG Cornell voltage (mV)	0.324	0.070	0.187–0.462	<0.0001
Model 2 (R^2 = 0.61)				
Age (years)	0.610	0.070	0.473–0.748	<0.0001
Glucose (mg/dl)	0.110	0.027	0.057–0.163	<0.0001
ECG Cornell voltage (mV)	0.295	0.102	0.095–0.495	0.004
Neutrophil counts ($\times 10^3/\mu l$)	2.630	0.478	1.690–3.569	<0.0001

Model 2 suggests an independent association between pulse pressure and neutrophil count after the significant influence of some confounders.

Linear regression is used with continuous (such as BP) outcomes, while logistic regression is used with binary outcomes (e.g. LV hypertrophy, yes vs. no). Proportional hazards (Cox) regression is used when the outcome is the elapsed time to an event (e.g. time from baseline evaluation to stroke event).

Multivariable linear regression is a method used to model the linear relationship between a dependent variable and one or more independent variables. The dependent variable is sometimes also called the predictand, and the independent variables the predictors. The underlying assumption of multiple linear regression is that, as the independent variables increase or decrease, the mean value of the outcome increases or decreases in a linear fashion.

A regression analysis is based on least-squares and the model is fit such that the sum-of-squares of differences of observed and predicted values is minimized. In a multivariable model, the regression coefficient for each variable is estimated by fitting the model to the data and adjusting for other variables in the model. In other words, a multivariable regression analysis provides predictions based on the combined predictive effect of predictors.

To assess the power of a linear regression model to predict outcome, the adjusted R^2 may be reported. The value of R^2 ranges from 0 to 1 and, multiplied by 100, R^2 can be thought of as the percentage of the variance in the outcome accounted for by the independent variables. In a model with an R^2 close to 1, the dependent variables together accurately predict outcome.

A recent analysis on postmenopausal hypertension from our group illustrated how to perform a multivariable linear regression analysis [7]. We tested the independent association of neutrophil count, a marker of chronic inflammation, with pulse pressure (PP), a recognized marker of atherosclerosis and risk factor for cardiac and cerebrovascular events.

Since the linear combination of age, serum glucose levels and ECG Cornell voltage was a good predictor of PP in a multivariable model (model 1; table 1), the independent association of neutrophil count with PP was tested after adjusting for the influence of these factors. Notably, in this multivariable model (model 2; table 1), the association between PP and neutrophil count remained significant after adjusting for the significant influence of these confounders.

Binary logistic regression estimates the probability of an outcome and models how that probability changes with a change in the predictor variables. The basic assumption is that each one-unit increase in a predictor multiplies the odds of the outcome by a certain factor and that the effect of several variables is the multiplicative product of their individual effects. The logistic function produces a probability of outcome bounded by 0 and 1. The central mathematical concept that underlines logistic regression is the logit (the natural logarithm of an odds ratio). The simplest example of a logit derives from a 2 × 2 contingency table.

Consider the same data from the PIUMA study, which assessed the association of LV hypertrophy with the risk of stroke. The distribution of a binary outcome variable

Table 2. Independent predictors of cerebrovascular events in the PIUMA study [data from ref. 4]

Variable	Units/comparison	HR (95% CI)	p value
Age	10 years	2.36 (1.63–3.43)	0.001
24-hour systolic BP	12 mm Hg	1.68 (1.17–2.38)	0.004
Serial changes in LVH	absent (coded as 0)	1 (reference group)	
	present (coded as 1)	2.80 (1.18–6.69)	0.020

Absent = Subjects with persistently normal LV mass or regression of LV hypertrophy; present = subjects with lack of regression or new development of LV hypertrophy.

(stroke, yes vs. no) is paired with a dichotomous predictor variable (LV hypertrophy, yes vs. no). Results suggest that patients with LV hypertrophy at baseline evaluation are 2.58 times more likely, than not, to develop a stroke compared to patients without LV hypertrophy. The odds ratio is derived from two odds and its natural logarithm is a logit, which equals 0.95. The value of 0.95 is the regression coefficient of the logistic regression. The antilogarithm of the regression coefficient equals the odds ratio for a one-unit increase in the predictor. In case of continuous explanatory variables, the units of change can be specified (e.g. 10-mm Hg increase in BP) for which the odds ratio is estimated.

Since systolic BP is a confounder for its association with both LV hypertrophy and cerebrovascular disease, we model a multivariable logistic regression including LV hypertrophy and systolic BP as predictors. After adjustment for the significant influence of systolic BP, the presence of LV hypertrophy is still associated to an increased risk of stroke (OR 1.98, 95% CI 1.37–2.86; $p < 0.0001$).

Proportional hazards models assume that the ratio of the hazards for subjects with and without a given risk factor is the constant over the entire study period [8]. This is known as the proportionality assumption and it is the primary concern when fitting a Cox model. This assumption implies that the survivor functions do not cross and explanatory variables act only on the hazard ratio. An advantage of proportional hazards analysis is that it includes subjects with varying lengths of follow-up. A subject who does not experience the outcome of interest by the end of the study is considered censored. The antilogarithm of the proportional hazards regression coefficient equals the relative hazard for a one-unit increase in the predictor. In case of continuous explanatory variables most modern software allows to specify the units of change (e.g. 10-mm Hg increase in BP) for which the customized hazard ratio is estimated.

Take, for example, a study of the association between regression of LV hypertrophy and risk of stroke [4]. Person-time analysis demonstrated that hypertensive subjects with lack of regression or new development of LV hypertrophy had a markedly increased rate of stroke when compared with subjects who never developed LV hypertrophy or with regression of LV hypertrophy (1.16 vs. 0.25 × 100 patients per year; $p = 0.0001$).

The independent effect of serial changes in LV hypertrophy was tested by multivariable Cox model. Other tested confounders were ambulatory 24-hour systolic BP, age, gender (men, women), body mass index, diabetes (no, yes), total cholesterol, serum triglycerides, family history of cardiovascular disease (no, yes), smoking habits, type of antihypertensive treatment at the follow-up visit, and statin treatment at the follow-up visit. In the multivariable analysis, the risk of cerebrovascular events was 2.8 times higher (95% CI 1.18–6.69) in the subset with lack of regression or new development of LV hypertrophy than in that with LV hypertrophy regression or persistently normal LV mass. Such an effect was independent of age and 24-hour systolic BP at the follow-up visit (table 2).

Cox's semiparametric model, because it does not assume any distribution for the baseline hazard [9], also allows time-dependent explanatory variables. An explanatory variable is time-dependent if its value changes over time. For instance, you can use a time-dependent variable to model the effect of subjects changing treatment groups or exposure status. Or you can include time-dependent variables such as BP that vary with time during the course of a study [10].

How Many Variables in a Model?

A common problem in regression analysis is variable selection. When examining the effect of a risk factor, different adjustments for other factors may yield different or even confusing conclusions. A variable selection method is a way of selecting a particular set of predictors for use in a regression model, or it might be an attempt to find a 'best' model when there are several candidate predictors. The decision on what to adjust should be guided by an a priori theoretical or biological relationship among the different factors and the outcome [11]. Conversely, the

number of variables in a model is often obtained by 'bivariate screening' or by using automated variable selection procedures such as forward, backward, or stepwise selection. Bivariate screening starts by looking at all bivariate relationships with the dependent variable, and includes any that are significant in a main model. Unfortunately, this is usually inadequate. Because of correlations among the explanatory variables, any one variable may have little unique predictive power, especially when the number of predictors is large. Automated procedures determine the order in which the predictor variables are entered into the model according to statistical criteria. In forward selection, variables are entered into the model one at a time in an order determined by the strength of their association with the criterion variable. The effect of adding each is assessed as it is entered, and variables that do not significantly add to the success of the model are excluded. In backward selection, all the predictor variables are entered into the model. The weakest predictor variable is then removed and the regression recalculated. If this significantly weakens the model then the predictor variable is re-entered, otherwise it is deleted. This procedure is then repeated until only useful predictor variables remain in the model. Stepwise selection alternates between forward and backward, bringing in and removing variables that meet the statistical criteria for entry or removal, until a stable set of variables is attained. If adding the variable contributes to the model then it is retained, but all other variables in the model are then retested to see if they are still contributing to the success of the model. If they no longer contribute significantly they are removed. This method theoretically defines the smallest possible set of predictor variables included in the final model.

Results from stepwise regression are sensitive to violations of the assumptions underlying regression. More generally, indiscriminate use of variable selection procedures might result in models with biased selection of variables, unreliable coefficients and inaccurate prediction.

Although there is little consensus on the best methods for selecting variables, the use of some methods is generally discouraged, such as inclusion or exclusion of variables based on univariable analysis. We suggest simple rules of thumb for selecting explanatory variables: include an adequate number of predictors to make the model useful for theoretical and practical purposes and to obtain good predictive power. Do not exclude variables solely in view of a nominally nonsignificant association or because they are, possibly by chance, not predictive in the particular sample. Adding variables with little predictive power has disadvantages. Redundant variables usually fail to improve model fit values because they do not add to the overall prediction. To prevent collinearity, it is helpful for the explanatory variables to be correlated with the response variable but not highly correlated between them.

The 'stopping rule' for inclusion or exclusion of predictors is a burning question in model selection. Besides standard significance level for testing of hypotheses ($\alpha = 0.05$), the use of Akaike's Information Criterion (AIC) and the Bayesian Information Criterion (BIC) is also very popular. AIC and BIC compare models based on their fit to the data, but penalize for model complexity, i.e. the number of degrees of freedom. AIC requires that the increase in model χ^2 has to be larger than two times the degrees of freedom. For instance, considering a predictor with 1 degree of freedom, such as gender, this implies that the model χ^2 has to exceed 2. BIC penalizes the model fit such that the model χ^2 has to exceed the number of predictors multiplied by logarithm of the effective sample size, e.g. the number of events in a Cox survival model. Models with lower AIC and/or BIC are usually preferred.

Finally, some hints can be given regarding the number of candidate predictors that can be reliably studied in relation to the sample size. A well-known rule of thumb is the 1 in 10 or 1 in 20 rule [12, 13]. For linear models, such rule suggests that 1 candidate predictor can be studied for every 10 or 20 patients. For logistic or Cox models, the 1 in 10 rule is rather superficial unless there is a fully prespecified set of predictors. Besides, it must be remembered that power and validity of a multivariable survival analysis is related to the number of outcome events compared to number of candidate predictors (i.e. the effective sample size) rather than the number of participants (total sample size). We suggest the 1 in 20 rule for these models with a limited set of prespecified predictors and the 1 in 50 rule for stepwise selection. Hence, in a study with 60 patients experiencing an outcome event (60 events) out of 3,000 exposed, only 3 prespecified predictors could reliably be studied according to the 1 in 20 rule. When the rule is violated, the number of candidate predictors is generally too large for the data set, and overfitting will almost inevitably occur.

Overfitting

What is overfitting? The principle of parsimony or Occam's razor dictates using models that contain all that is necessary for the modeling but nothing more. If a simpler

model is statistically indistinguishable from a more complex model, parsimony dictates that we should prefer the simpler model. For example, if a regression model with 3 predictors is enough to explain the outcome, then no more than these predictors should be used. Moreover, if the relationship can be captured by a linear function in these predictors, then using a quadratic term violates parsimony. Overfitting is the use of models or procedures that violate parsimony, that is, that include more terms than are necessary or use more complicated approaches than are necessary.

Goodness of Fit

One key aspect of multivariable regression modeling is how well the model agrees with the data, i.e. the goodness of fit of the model. Knowledgeable authors pointed out that although goodness of fit is fundamental to evaluate the validity of regression models, it is sparsely reported in published articles [14, 15]. For instance, the goodness of fit of logistic models is usually evaluated as follows: first, use global measures of model fit, such as likelihood statistics, and, second, evaluate individual observations to see whether any are problematic for the regression model. Residual analysis is an effective way to detect outliers or overly influential observations. Large residuals suggest that the model does not fit the data. Unfortunately, medical journal articles rarely, if ever, present residual plots.

Interactions

According to Concato et al. [16], 'An interaction occurs between independent variables if the impact of one variable on the outcome event depends on the level of another variable.' Multivariable methods do not automatically assess interactions, which can be evaluated by explicitly adding interaction terms to the model. When an interaction effect is present, the impact of one variable depends on the level of the other variable and interpretation might not be straightforward. For instance, an intervention study tests the effects of a treatment on an outcome measure. The treatment variable is composed of two groups, treatment and control. The outcome incidence in the treatment group is lower than in the control group. However, from previous studies we hypothesize that treatment effect may not be equal for men and women, i.e. is there a difference in treatment depending on gender? This is a question of interaction, and to address it we would add a specific interaction term (treatment by gender) to the model. However, the inclusion of interactions, when the study was not specifically designed to assess them, can make it difficult to estimate and interpret the other effects in the model. Hence, if a study was not specifically designed to assess interactions and there is no a priori reason to expect one, or interaction terms are being assessed just because statistical software makes it simple, and no interaction is actually found, it might be wise to fit the model without the interaction term given the absence of a universal rule dictating appropriate tests for interactions in all circumstances [16].

Conclusions

Our purpose was to introduce clinical readers, often uncomfortable with statistics, to multivariable analysis using practical suggestions and nontechnical language. In particular, we reviewed the basics of multivariable models most commonly used in clinical research, how they are assembled, and how they can be interpreted and evaluated.

References

1 Katz MH: Multivariable Analysis: A Practical Guide for Clinicians and Public Health Researchers. London, Cambridge University Press, 2011.
2 Whisnant JP: Modeling of risk factors for ischemic stroke. The Willis Lecture. Stroke 1997;28:1840–1844.
3 Verdecchia P, Angeli F, Achilli P, Castellani C, Broccatelli A, Gattobigio R, et al: Echocardiographic left ventricular hypertrophy in hypertension: marker for future events or mediator of events? Curr Opin Cardiol 2007; 22:329–334.
4 Verdecchia P, Angeli F, Gattobigio R, Sardone M, Pede S, Reboldi GP: Regression of left ventricular hypertrophy and prevention of stroke in hypertensive subjects. Am J Hypertens 2006;19:493–499.
5 Verdecchia P, Angeli F, Gattobigio R, Guerrieri M, Benemio G, Porcellati C: Does the reduction in systolic blood pressure alone explain the regression of left ventricular hypertrophy? J Hum Hypertens 2004;18(suppl 2):S23–S28.
6 Katz MH: Multivariable analysis: a primer for readers of medical research. Ann Intern Med 2003;138:644–650.
7 Angeli F, Angeli E, Ambrosio G, Mazzotta G, Cavallini C, Reboldi G, et al: Neutrophil count and ambulatory pulse pressure as predictors of cardiovascular adverse events in postmenopausal women with hypertension. Am J Hypertens 2011;24:591–598.

8 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;15:361–387.
9 Rao SR, Schoenfeld DA: Survival methods. Circulation 2007;115:109–113.
10 Kleinbaum DG, Klein M: Survival Analysis: A Self-Learning Text, ed 2. New York, Springer, 2005.
11 Greenland S: Modeling and variable selection in epidemiologic analysis. Am J Public Health 1989;79:340–349.
12 Harrell FE: Regression Modeling Strategies: with Applications to Linear Models, Logistic Regression, and Survival Analysis. New York, Springer, 2001.
13 Steyerberg EW: Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating. New York, Springer, 2009.
14 LaValley MP: Logistic regression. Circulation 2008;117:2395–2399.
15 Bagley SC, White H, Golomb BA: Logistic regression in the medical literature: standards for use and reporting, with particular attention to one medical domain. J Clin Epidemiol 2001;54:979–985.
16 Concato J, Feinstein AR, Holford TR: The risk of determining risk with multivariable models. Ann Intern Med 1993;118:201–210.