Published in final edited form as:
Am J Public Health. 2013 January ; 103(1): . doi:10.2105/AJPH.2012.300897.

Multivariate or Multivariable Regression?

Bertha Hidalgo, PhD, MPH and
Department of Biostatistics, Section on Statistical Genetics, University of Alabama at Birmingham

Melody Goodman, PhD, MS
Department of Surgery, Division of Public Health Sciences, School of Medicine, Washington University in St. Louis, St. Louis, MO

Abstract

The terms multivariate and multivariable are often used interchangeably in the public health literature. However, these terms actually represent 2 very distinct types of analyses. We define the 2 types of analysis and assess the prevalence of use of the statistical term multivariate in a 1-year span of articles published in the American Journal of Public Health. Our goal is to make a clear distinction and to identify the nuances that make these types of analyses so distinct from one another.

Most regression models are described in terms of the way the outcome variable is modeled: in linear regression the outcome is continuous, logistic regression has a dichotomous outcome, and survival analysis involves a time to event outcome. Statistically speaking, multivariate analysis refers to statistical models that have 2 or more dependent or outcome variables,[1] and multivariable analysis refers to statistical models in which there are multiple independent or response variables.[2]

A multivariable model can be thought of as a model in which multiple variables are found on the right side of the model equation. This type of statistical model can be used to attempt to assess the relationship between a number of variables; one can assess independent relationships while adjusting for potential confounders.

A simple linear regression model has a continuous outcome and one predictor, whereas a multiple or multivariable linear regression model has a continuous outcome and multiple predictors (continuous or categorical). A simple linear regression model would have the form

$$y=\alpha+x\beta+\varepsilon \quad (1)$$

By contrast, a multivariable or multiple linear regression model would take the form

$$y=\alpha+x_1\beta_1+x_2\beta_2+\ldots+x_k\beta_k+\varepsilon \quad (2)$$

Correspondence should be sent to Bertha Hidalgo, PhD, MPH, 1665 University Blvd, RPHB 443, Birmingham, AL 35294 (bhidalgo@uab.edu).

Reprints can be ordered at http://www.ajph.org by clicking the “Reprints” link.

Contributors
Both authors contributed equally to this article. B. Hidalgo conducted the literature review and led the writing. M. Goodman conceived the topic and supervised the development of the article.

Human Participant Protection
No protocol approval was needed because no human subjects were involved.

where y is a continuous dependent variable, x is a single predictor in the simple regression model, and x_1, x_2, …, x_k are the predictors in the multivariable model.

As is the case with linear models, logistic and proportional hazards regression models can be simple or multivariable. Each of these model structures has a single outcome variable and 1 or more independent or predictor variables.

Multivariate, by contrast, refers to the modeling of data that are often derived from longitudinal studies, wherein an outcome is measured for the same individual at multiple time points (repeated measures), or the modeling of nested/clustered data, wherein there are multiple individuals in each cluster. A multivariate linear regression model would have the form

$$\mathrm{Y}_{\mathrm{n}\times\mathrm{p}}=\mathrm{X}_{\mathrm{n}\times(\mathrm{k}+1)}\beta_{(\mathrm{k}+1)\times\mathrm{p}}+\varepsilon \quad (3)$$

where the relationships between multiple dependent variables (i.e., Ys)—measures of multiple outcomes—and a single set of predictor variables (i.e., Xs) are assessed.

We took a systematic approach to assessing the prevalence of use of the statistical term multivariate. That is, we used PubMed and the keyword "multivariate" to review articles published in the *American Journal of Public Health* over a 1-year span (December 2010–November 2011). We identified 30 articles in which the authors indicated the use of a "multivariate" statistical method. Each of the articles was individually reviewed to assess the type of analysis defined as multivariate.

In 5 (17%) of the 30 articles, multivariate models (as we have defined them here) were used; 4 (13%) of these models were derived from longitudinal data and 1 from nested data. The remaining 25 (83%) articles involved multivariable analyses; logistic regression (21 of 30, or 70%) was the most prominent type of analysis used, followed by linear regression (3 of 30, or 10%). Interestingly, in 2 of the 30 articles (7%), the terms multivariate and multivariable were used interchangeably. This further elucidates the need to establish consistency in use of the 2 statistical terms.

Although some may argue that the interchangeable use of multivariate and multivariable is simply semantics, we believe that differentiating between the 2 terms is important for the field of public health. In general, models used in public health research should be described as simple or multivariable, to indicate the number of predictors, and as linear, logistic, multivariate, or proportional hazards, to indicate the type of outcome (e.g., continuous, dichotomous, repeated measures, time to event).

Our review revealed that there is a need for more accurate application and reporting of multivariable methods. This issue is not unique to public health research and has been identified as affecting other areas of research as well (e.g., medicine, psychology, political science).[3] However, we hope to see a clearer distinction in the use of the terms multivariate and multivariable to describe statistical analyses in future public health literature. This is an important distinction not only to avoid confusion among readers but to more accurately inform the next generation of public health researchers who are seeking to ground their work in the published literature.

Acknowledgments

B. Hidalgo was supported in part by a predoctoral training grant from the National Cancer Institute (grant 5R25CA047888) and a postdoctoral training grant from the National Heart, Lung, and Blood Institute (grant T32HL072757). M. Goodman was supported by the Siteman Cancer Center, the National Cancer Institute (grant U54CA153460), and the Washington University Faculty Diversity Scholars Program.

References

1. Van Belle, G. Biostatistics: A Methodology for the Health Sciences. Hoboken, NJ: Wiley-Interscience; 2004.
2. Katz MH. Multivariable analysis: a primer for readers of medical research. Ann Intern Med. 2003; 138(8):644–650. [PubMed: 12693887]
3. Freedland KE, Reese RL, Steinmeyer BC. Multivariable models in biobehavioral research. Psychosom Med. 2009; 71(2):205–216. [PubMed: 19218467]