Do behavioural ecologists misuse the term multivariate?

Clint D. Kelly

Département des Sciences biologiques

Université du Québec à Montréal

Montreal, QC H2X 1Y4

kelly.clint@uqam.ca

Abstract

Clear communication of procedures and findings is critical for scientific progress. One problem facing many research disciplines is the incorrect labelling in the literature of multivariable (i.e. many Xs) statistical models as multivariate (many Ys). Multivariate and multivariable statistical models are different statistical approaches and should be described as such. I assessed whether behavioural ecologists fall victim to the mislabelling plaguing other disciplines by reviewing relevant papers published in Behavioral Ecology between 2013-2023. My review shows that approximately 30% of multivariable models were incorrectly labelled as multivariate. Although this error rate is less severe than in other scientific disciplines it is nonetheless greater than zero.

Background

Scientific progress requires the precise definition and consistent use of terms in the literature. Inconsistent usage and misapplication of terms in our publications will confuse our audience, cause them to misinterpret our research findings, and ultimately retard progress. An example of terminological misuse that is observed in many scientific research fields is the incorrect use of the statistical terms multivariate and multivariable (Hidalgo & Goodman, 2013).

Multivariate analysis refers to statistical models that have two or more dependent or outcome variables (i.e. many Ys) (Hidalgo & Goodman, 2013). This type of analysis is increasingly being used by behavioural ecologists, for example, to analyze the repeated measures of behaviour to identify animal personalities (Dingemanse & Dochtermann, 2013) and use geometric morphometric approaches to analyze trait shape (Adams, Rohlf, & Slice, 2013). In contrast, multivariable analysis refers to statistical models in which there are multiple independent or explanatory variables (i.e. many Xs) (Freedland, Reese, & Steinmeyer, 2009; Katz, 2003). This type of analysis is also commonly used by behavioural ecologists, for example, when using multiple regression to conduct a Lande-Arnold (Lande & Arnold, 1983) selection analysis. These are distinct statistical approaches yet in many research disciplines these terms are used interchangeably, leading to confusion amongst readers (Hidalgo & Goodman, 2013; Wong et al., 2018). Errors regarding this nomenclature are so common in some fields, and the problem is deemed so significant, that journals are adopting guidelines to ensure that authors properly describe their statistical models (Peters, 2008). Given the wide use of multivariate and multivariable analyses in

behavioural ecology, I asked whether our field suffers from the terminological misuse observed in other disciplines, and if so, recommend some simple steps that authors can take to alleviate the confusion stemming from the misuse.

Methods

I systematically assessed how the statistical terms 'multivariate' and 'multivariable' were used in the behavioural ecology literature by conducting two separate searches of the keywords 'multivariate' and 'multivariable' in the full text (but not in the references or bibliography) of articles published in *Behavioural Ecology* from 2013 to 2023. The searches were conducted by using the 'advanced search' function on the *Behavioral Ecology* website (**https://academic.oup.com/beheco**). Each of the identified articles was individually reviewed to assess the type of analysis defined as multivariate or multivariable.

Results and Discussion

My search identified 134 articles using the word multivariate of which 129 were available as full text. Thirty articles did not use the word multivariate to describe a statistical model (e.g. the word was used to describe a concept in the Introduction or Discussion or used a multivariate approach but did not identify it as such in the Methods, i.e. PCA) and were eliminated from the final tally. Of the remaining 99 papers, 70 used the term multivariate correctly (i.e. more than one response variable) and 29 misused it to describe a multivariable model (i.e. more than one explanatory factor). Although the error rate observed here (30%) is seemingly high, it is more encouraging than the 83% of public health and epidemiology papers surveyed by Hidalgo and Goodman (2013) that incorrectly

described a multivariable model as multivariate. Only six papers used the word 'multivariable' and of these, one incorrectly labelled their multivariate analysis of variance (MANOVA) as a multivariable analysis of variance while the other five correctly used the term to describe a model having one response variable and several explanatory variables.

Table 1. Summary of *Behavioural Ecology* papers (N=99) published between 2013 and 2023 in which authors correctly or incorrectly described their statistical model as multivariate.

Model is…	**Model described as…**	**N**
Multivariate	Multivariate	70
Multivariable	Multivariate	29

My literature survey revealed that behavioural ecologists use the term multivariate to describe multivariable models and suggest a need for more accurate application and reporting of multivariable methods. A more precise use of the terms multivariate and multivariable to describe statistical analyses is not simply semantics but rather a productive step to avoid confusion amongst readers.

The term multivariate was coined in the early 20th century along with two others commonly used statistical terms: univariate and bivariate. The early 1900s were the nascence of statistical modelling when data analysis was relatively straightforward. Today, we still use these terms despite models being much more sophisticated and complex. Consequently, our

current usage of these terms does not always align with their original definition. For example, univariate describes instances in which there is only one response variable per observation (Wishart, 1928), thus allowing one to calculate measures of central tendency and frequency distribution but not associations or relationships between variables (Devore & Peck, 1986). My article survey, however, found that authors use the term 'univariate' to describe statistical models such as simple linear regression (univariate regression) and one-way analysis of variance (univariate ANOVA). The problem with using these terms is that it is not clear how many explanatory variables are in the model in contrast to using 'simple linear regression' or 'multiple regression', which provide some indication. Not only should we avoid making up new terms when existing ones are suitable (and well-established) but we should also avoid creating new names that are redundant. We assume that simple linear regression and one-way ANOVA have only one response variable so there is little need to state that it is univariate.

Pearson (1920) coined the term 'bivariate' to distinguish a simple correlation (i.e. a relationship between two variables) from the more advanced 'multivariate' correlation (i.e. a relationship between three or more variables) that was being developed at the time. Bivariate is still commonly used to describe relationships between two variables (e.g. Pearson correlation) as well as models having two response variables; however, my literature survey revealed that it is also used to describe all manner of models including regressions having one response variable and one or two explanatory factors rather than using the well-established terms simple linear regression or multiple linear regression, respectively. If new monikers are required for these well-established terms, then I suggest

that we refer to them from the perspective of their explanatory factors (i.e. univariable and multivariable regressions) and not the number of response variables, which is assumed to be one. We should continue to use 'bivariate' to refer to cases in which only two variables are analyzed (Tsai, 2013) while reserving 'multivariate' for models having more than one response variable and a single explanatory variable. Cases in which there are multiple response and explanatory variables should be called "multivariate-multivariable".

In summary, I recommend that we simplify and clarify our language when describing statistical models by categorizing a model as either univariable, multivariable, multivariate, or multivariate-multivariable. When the model has a single response variable, we should describe it from the perspective of the number of explanatory variables (i.e. univariable or multivariable). If the model has more than one response variable and only one explanatory variable, then we should describe it based on the number of response variables (i.e. multivariate). If the model has many response and explanatory variables, then it should be called a multivariate-multivariable model. Finally, one way to eliminate any confusion is to include the model's formula in the manuscript, a practice gaining popularity in ecology and evolution journals. This step will certainly help clarify how the data were analysed but it will have a high barrier to entry as most of us are much better at verbally describing our models than notating them. However, the cost of this approach is that incorrect model formulation will compound our confusion by creating a discrepancy between the verbal and algebraic descriptions. Model equations notwithstanding, the simple terminological changes that I suggest here will reduce confusion amongst readers and permit better appraisal of our statistical methodology and research findings. Even if readers disagree with my

suggestions, I hope that this commentary at least makes behavioural ecologists think critically and clearly about how they describe their models.

References

Adams, D. C., Rohlf, F. J., & Slice, D. E. (2013). A field comes of age: Geometric morphometrics in the 21st century. *Hystrix*, *24*(1), 7-14.

Devore, J. L., & Peck, R. (1986). *Statistics, the Exploration and Analysis of Data*. West Group.

Dingemanse, N. J., & Dochtermann, N. A. (2013). Quantifying individual variation in behaviour: mixed-effect modelling approaches. *Journal of Animal Ecology*, *82*(1), 39-54.

Freedland, K. E., Reese, R. L., & Steinmeyer, B. C. (2009). Multivariable models in biobehavioral research. *Psychosomatic Medicine*, *71*(2), 205-216.

Hidalgo, B., & Goodman, M. (2013). Multivariate or multivariable regression. *American Journal of Public Health*, *103*(1), 39-40.

Katz, M. H. (2003). Multivariable Analysis: A Primer for Readers of Medical Research. *Annals of Internal Medicine*, *138*(8), 644.

Lande, R., & Arnold, S. J. (1983). The measurement of selection on correlated characters. *Evolution*, *37*(6), 1210-1226.

Pearson, K. (1920). Notes on the History of Correlation. *Biometrika*, *13*(1), 25-45.

Peters, T. J. (2008). [C] Multifarious terminology: multivariable or multivariate? univariable or univariate. *Paediatric and perinatal epidemiology*.

Tsai, A. C. (2013). Achieving Consensus on Terminology Describing Multivariable Analyses. *American Journal of Public Health*, *103*(6), e1-e1.

Wishart, J. (1928). The generalised product moment distribution in samples from a normal multivariate population. *Biometrika*, *20A*, 32-52.

Wong, E., Oh, L. J., Andrici, J., McCluskey, P., Smith, J. E. H., & Gill, A. J. (2018). Author reply. *Intern Med J*, *48*(5), 608.