Proper Analysis of Categorical Outcomes: Make Sure You Understand the Model

Andrew D Althouse[*]

University of Pittsburgh, Medical Center Pittsburgh, PA, USA

[*]**Corresponding author:** Andrew D. Althouse, University of Pittsburgh, Medical Center Pittsburgh, PA, USA, Tel: +14843328876; E-mail: althousead@upmc.edu

Introduction

This letter is written in response to the recent publication by Darabont et al. [1] regarding the possible relationship between Acute Pulmonary Edema (APE) and Renal Artery Stenosis (RAS). In this publication, the authors used a statistical technique known as "linear discriminant analysis" to assess the relationships between selected predictor variables (including APE) and their study outcome (RAS). Although the authors are commended for taking on this investigation, the choice of statistical analysis is inappropriate for their data, and has some technical assumptions which make it unsuitable for the manner in which it was used by Darabont et al.

Briefly, linear discriminant analysis is meant to find a linear combination of factors that correctly predict or characterize a certain event. This may sound like technical jargon, but simply put it is a way to predict a categorical outcome variable using continuous predictor variables. To the statistically fluent reader, this may sound a lot like logistic regression, and it should; logistic regression, also, essentially creates a model using a linear combination of factors to predict a categorical outcome variable. However, there are key technical differences between the two, and there is a reason that logistic regression remains highly prevalent today while linear discriminant analysis is rarely seen.

In 1978, Press and Wilson [2] compared linear discriminant analysis to logistic regression and found that logistic regression was the superior technique in the vast majority of cases. One principal reason is that linear discriminant analysis has some highly specific assumptions which are almost never met; specifically, that the underlying variables follow a jointly normal distribution. This cannot be true in the case of Darabont et al, because several of the predictor variables used in their model is most assuredly not normally distributed (including gender, which is a dichotomous variable and thus cannot possibly follow a normal distribution). However, logistic regression can be used for the purpose which they desired, namely creating a multivariate model using both continuous and categorical variables to predict a binary outcome (such as RAS).

This does not entirely invalidate the results published by Darabont et al; it is possible, even likely, that replication of their efforts using logistic regression would have yielded similar conclusions. However, it is disturbing to this reader to see published work that includes such rudimentary mistakes in the statistical analysis. Logistic regression is a basic statistical technique that can be performed in any statistical software package, including freely downloadable software such as R [3], and is common knowledge to any graduate-level statistics student. In the future, authors submitting to the Journal should examine their analytic choice with caution to ensure that they are using proper statistical methods for their study question. If they are uncertain, they should consult a professional statistician to ensure that their choice of technique is appropriate.

References

1. Darabont RO, Corlan AD, Vinereanu D (2015) Renal artery stenosis and acute pulmonary edema: a possible correlation beyond Pickering Syndrome. J Clin Exp Cardiolog 6: 6.

2. Press SJ, Wilson S (1978) Choosing between logistic regression and discriminant analysis. Journal of the American Statistical Association 364: 699-705.

3. https://cran.r-project.org/