Journal of Biometrics & Biostatistics
ISSN: 2155-6180
Thornley and Marshall, J Biomet Biostat 2012, 3:6
DOI: 10.4172/2155-6180.1000e111
Editorial Open Access

Measures of Association in Epidemiological Studies: How Best to Compare Discrete and Continuous Variables?

Simon Thornley* **and Roger Marshall**

Section of Epidemiology and Biostatistics, The University of Auckland, Tamaki Campus, Glen Innes, 1142, Auckland, New Zealand

The strength of the association in an observational study, often measured by a relative risk estimate, is an important factor in determining whether an exposure is likely to be causally related to the outcome. One Bradford-Hill criterion is that strength of association shifts the weight of evidence toward causation [1]. Assessing the relative strength of association between variables in a multivariate model is thus an important part of epidemiological data analysis.

Today many variables may find their way into regression analyses. They may include both discrete (categorical) and continuous variables and be measured on non-commensurate scales. Direct comparison of beta-coefficients or measures of association derived from them, is often not meaningful.

One approach to this problem is to simply categorise continuous data, making all predictors binary, thus resolving the problem of comparison. While this can render the results of analyses straightforward to interpret, objections exist because it reduces precision and statistical power and assumes the relationship between exposure and outcome is the same within intervals. It may also introduce bias, if the categorisation is made by the analyst who is not blinded to the output of his or her analyses. This approach also does not condition on all the exposure information available, and so is likely to result in residual confounding.

With the availability of spline functions and polynomial terms in regression models, the assumption of monotonic and linear relationships between exposure and outcome may be overcome; however, such techniques present difficulties of interpretation.

One solution to express the strength of an association is to report standardised regression coefficients. This simply involves multiplying estimated beta coefficients by the standard deviation of the variable, so that the beta coefficient of interest is transformed to the same scale as all others and the standardised coefficient expresses the effect for a unit standard deviation shift in each variable. This has a disadvantage, if consistently applied to both binary and continuous variables, because it means that the "strength" of a binary variable depends on the prevalence of the variable p since the standardized coefficient is β times $\sqrt{p(1-p)}$.

For example, if the beta coefficients of x_1 and x_2 are both β=1 say, but x_1 and x_2 have proportions p_1=0.5 and p_2=0.1, the standardised coefficients are 0.5 and 0.3 suggesting x_1 to have a stronger association on the outcome than x_2. But whether prevalence of a variable should determine a variable's "strength" of association seems arguable. Further, there seems little to be gained by reporting a standardized β of 0.5 over its actual value of β=1.

For binary variables, therefore it seems preferable not to standardise. However, for continuous measures, there is some justification in reporting coefficients that are expressed with respect to commensurate scales and, further, to ensure that effects can be directly compared to binary effects. Standardising with respect to standard deviation is one idea [2], another is to consider the effect relative to upper and lower values of the variable. This feature for example, in Frank Harrell's *rms* regression package [3] for R in which the upper and lower quartiles are by default, the reference points. For a binary variable, 0 is the lower and 1 the upper value, which is a range of two standard deviations of a binary variable with p=0.5. For normally distributed variables, the distance between the 16th and 84th percentiles is also approximately two standard deviations so that, taking the binary case as a benchmark, we suggest that the 16th and 84th percentiles make more appropriate reference values than quartiles.

As an example, in a recent analysis, not yet published, of the effect of serum urate on incident cardiovascular disease, in New Zealand adults, the adjusted hazard ratio of serum urate, comparing the 16th and 84th (0.27 and 0.45 mmol/L) centiles, was 1.56 (95% confidence interval: 1.32 to 1.84) [4]. The strength of such an association may be difficult to grasp in isolation, as other studies use different comparisons, and cutoffs (such as quintiles, or binary variables, with varying definitions of hyperuricemia). With the knowledge that the distributional change is similar to a binary variable, such as smoking (compared with non-smokers), for example (HR 1.63; 95% CI: 1.43 to 1.86), the measure of association for urate may be interpreted as of similar magnitude to this common health-influencing behaviour.

Though this idea is simple, we believe that it is helpful to communicate the relative strength of association in regression models.

References

1. Hofler M (2005) The Bradford Hill considerations on causality: a counterfactual perspective. Emerging Themes in Epidemiology 2: 11.
2. Schlesselman JJ (1982) Case-control studies: design, conduct, analysis. Oxford University Press, Oxford, United Kingdom.
3. Frank EH (2011) Regression Modeling Strategies 3.3-2 version, Springer-Verlag, USA.
4. Thornley S, Marshall R, Jackson R, Gentles D, Dalbeth N, et al. (2012) Is serum urate causally associated with incident cardiovascular disease? University of Auckland, Auckland, New Zealand.

***Corresponding author:** Simon Thornley, Section of Epidemiology and Biostatistics, The University of Auckland, Tamaki Campus, Glen Innes, 1142, Auckland, New Zealand, Tel: +6493737599 ext 81971; Fax: +64 9 3737503, E-mail: s.thornley@auckland.ac.nz

Received May 16, 2012; **Accepted** May 16, 2012; **Published** May 18, 2012

Citation: Thornley S, Marshall R (2012) Over Measures of Association in Epidemiological Studies: How Best to Compare Discrete and Continuous Variables? J Biomet Biostat 3:e111. doi:10.4172/2155-6180.1000e111