RHEUMATOLOGY

Rheumatology 2015;54:1133–1134
doi:10.1093/rheumatology/kev116
Advance Access publication 13 May 2015

Editorial

Statistical modelling: essentially, all models are wrong, but some are useful

Review series on statistical modelling

This editorial refers to Introduction to statistical modelling: linear regression, by Mark Lunt, on pages 1137–40 and Introduction to Statistical modelling 2: categorical variables and interactions in linear regression, by Mark Lunt, on pages 1141–44.

In rheumatology research, we are often interested in the association between multiple factors and an outcome. For example, with the current focus on individualized health care, we are interested in predicting the outcome of a specific treatment based on a set of variables. In this situation, basic statistical methods based on stratifying data on (co-)variables are often no longer sufficient because with multiple predictors, strata become too small to obtain meaningful results and statistical modelling becomes necessary.

The advantage of statistical modelling is that it results in a regression model that compactly represents the analysis results (i.e. as one function). However, it comes at the cost of extra assumptions. Another disadvantage is that the models used or the interpretation of the regression coefficients may not be well understood by the intended audience or even the user [1].

In this era of computationally powerful, widely available and user-friendly statistical software, statistical modelling is now a possibility for many of us without the necessary input from a statistician or methodologist. In fact, access to a statistician might be more limited than access to statistical software in many situations. Training of (clinical) researchers is therefore imperative.

In this issue of *Rheumatology*, Mark Lunt contributes to this training by reviewing an often-used modelling technique: linear regression [2]. Given that in rheumatology our outcomes are often continuous in nature (although we sometimes categorize them), this is often an appropriate statistical model. The focus of his discussion is, appropriately, on the practical application of linear regression. Importantly, as assumptions need to be satisfied for valid modelling results, Lunt also discusses how to verify these assumptions as a vital part of the analysis, which is probably often neglected [1].

Moreover, a model is always a simplification of the truth because it is unlikely that all variables related to the outcome, including variables that can confound the association under study (confounders), are included in the analysis or even measured in the study and that the actual (causal) relations follow the linear model perfectly. Hence, the famous remark by statistician, George E. P. Box: 'Essentially, all models are wrong, but some are useful' [3].

Our models should result in clinically useful prediction or improved understanding of how factors relate to each other. Therefore, an understanding of the regression function (as the main result of the analysis) is crucial. In the review by Lunt, the interpretation of this regression function is clearly explained using an example from the Journal.

More generally, regression models should be tailored by the researcher to suit the specific question. For questions regarding prediction, prognosis or diagnosis, or a question regarding causality, that is, related to the effect of treatment or an aetiological factor, modelling can all be relevant; however, the specific modelling strategy might be different.

Even when using statistical modelling, the total number of possible variables and possibilities in make-up of our models is often too large to test them all in our data sets. For instance, a variable may be used in the model as continuous but also as categorical (i.e. a grouping variable). It might also be of interest to test whether the effect of a variable is different between categories of another variable (i.e. interaction or effect modification). Both issues will be discussed in the second review of this series [4]. Decisions about which variables to evaluate in the modelling process should therefore be taken actively by the researcher and guided by clinical expertise. They should usually not be left to mechanical algorithms as often present in statistical software (i.e. forward or backward variable selection strategies), which are usually based purely on statistical significance [5, 6].

Lunt also touches on the distinction between statistical significance testing and estimation using 95% CIs, the latter being more informative. Furthermore, when we are interested in a combination of variables that best predicts the outcome, formal statistical significance of all variables in the model might be less important [6]. If we simply add a variable to a model to adjust for confounding, we would probably leave this factor in even if it is not statistically significant; it might still influence the (regression coefficient of the) association of interest. In this case also, we are guided by the specific research question and it is also related to the sample size needed. When developing the design of a study and building a model for analysis, graphical presentation of the variables and how they are (hypothesized to be) related to each other and the outcome can therefore be helpful [7].

Thus, developing a good model requires familiarity with the prerequisites and interpretation of statistical models as well as with the clinical subject matter. Previous knowledge about the question at hand and expert opinion should be used when developing such a model. In this regard, statistical modelling is an art rather than only a science, just like medicine.

Importantly, even if models are built well, there is always the risk of over-fitting them, so that the model explains the observed data well but performs much less well in a new but similar situation. Before applying the models in clinical practice, for example in prognostication or treatment selection, they should always be validated [8, 9]. Not only should internal validation, for instance using bootstrapping, be done as a minimum, but also external validation should be done by applying the prediction score (based on the regression function) to new (similar) data and then comparing predicted outcomes with observed outcomes.

As a clinical research community, we are obliged to use statistical techniques optimally to answer our research questions as we use data collected from patients who volunteer to participate in research. As stated by the Ethical Guidelines for Statistical Practice of the American Statistical Association, the person doing the analysis (not just the statistician) should have a competent understanding of the subject matter and use sound statistical methodologies suitable to the data [10]. This requires a basic understanding of the statistical techniques used at a level where it is at least known when it is important to consult a methodologist.

In conclusion, as a clinical epidemiologist, I highly support this series and hope that there will be more papers about statistical techniques that are useful for our research as, for instance, techniques to model categorical outcomes. I hope this and the forthcoming paper will be kept in files of clinical researchers for future reference so that the statistical models we develop will be clinically useful.

Funding: No specific funding was received from any funding bodies in the public, commercial or not-for-profit sectors to carry out the work described in this manuscript.

Disclosure statement: The author has declared no conflicts of interest.

Paco M. J. Welsing[1]

[1]*Department of Rheumatology and Immunology, University Medical Centre Utrecht, Utrecht, The Netherlands*

Revised version accepted 12 March 2015

Correspondence to: Paco M. J. Welsing, Department of Rheumatology and Immunology, University Medical Centre Utrecht, PO Box 85500, 3508 GA, Utrecht, The Netherlands. E-mail: paco.welsing@umcutrecht.nl

References

1 Greenland S. Introduction to regression models. In: Rothman KJ, Greenland S, Lash TL, eds. Modern epidemiology, 3rd edn. Philadelphia: Lippincott Williams & Wikins, 2008:381–417.

2 Lunt M. Introduction to statistical modelling: linear regression. Rheumatology 2015;54:1137–40.

3 Box GEP, Draper NR. Empirical model-building and response surfaces. New York, NY: Wiley, 1987: Vol. 424.

4 Lunt M. Introduction to statistical modelling 2: categorical variables and interactions in linear regression. Rheumatology 2015;54:1141–44.

5 Sauerbrei W, Royston P, Binder H. Selection of important variables and determination of functional form for continuous predictors in multivariable model building. Stat Med 2007;26:5512–28.

6 Ioannidis JPA, Greenland S, Hlatky MA *et al*. Research: increasing value, reducing waste 2. Increasing value and reducing waste in research design, conduct, and analysis. Lancet 2014;383:166–175.

7 Jupiter DC. Causal diagrams and multivariate analysis I: a quiver full of arrows. J Foot Ankle Surg 2014;53:672–3.

8 Collins GS, de Groot JA, Dutton S *et al*. External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. BMC Med Res Methodol 2014;14:40.

9 Moons KGM, Kengne AP, Grobbee DE *et al*. Review: risk prediction models: II. External validation, model updating, and impact assessment. *Heart* 2012;98:691–8.

10 Committee on Professional Ethics. 1999. American Statistical Association. Ethical guidelines for statistical practice. www.amstat.org/about/ethicalguidelines.cfm (7 January 2015, date last accessed).