Journal of Informetrics 8 (2014) 963–971

Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

Regression for citation data: An evaluation of different methods

Mike Thelwall*, Paul Wilson

Statistical Cybermetrics Research Group, School of Mathematics and Computer Science, University of Wolverhampton, Wulfruna Street, Wolverhampton WV1 1LY, UK

ARTICLE INFO

Article history:
Received 2 July 2014
Received in revised form
26 September 2014
Accepted 30 September 2014
Available online 22 October 2014

Keywords:
Informetrics
Altmetrics
Citation distributions
Lognormal
Powerlaw
Regression

ABSTRACT

Citations are increasingly used for research evaluations. It is therefore important to identify factors affecting citation scores that are unrelated to scholarly quality or usefulness so that these can be taken into account. Regression is the most powerful statistical technique to identify these factors and hence it is important to identify the best regression strategy for citation data. Citation counts tend to follow a discrete lognormal distribution and, in the absence of alternatives, have been investigated with negative binomial regression. Using simulated discrete lognormal data (continuous lognormal data rounded to the nearest integer) this article shows that a better strategy is to add one to the citations, take their log and then use the general linear (ordinary least squares) model for regression (e.g., multiple linear regression, ANOVA), or to use the generalised linear model without the log. Reasonable results can also be obtained if all the zero citations are discarded, the log is taken of the remaining citation counts and then the general linear model is used, or if the generalised linear model is used with the continuous lognormal distribution. Similar approaches are recommended for altmetric data, if it proves to be lognormally distributed.

1. Introduction

The use of performance monitoring for university research has increased over the past few decades. This is most evident in national research evaluation exercises, such of those in the UK (Mryglod, Kenna, Holovatch, & Berche, 2013), Australia (ARC, 2014), New Zealand (Anderson, Smart, & Tressler, 2013) and Italy (Abramo, D'Angelo, & Di Costa, 2011). This climate not only affects the allocation of research funding in many cases but can also change the behaviour of individual researchers as they come to terms with the assessment system (Butler, 2003). Although the most important performance monitoring exercises often rely on peer review, both the UK (REF, 2013) and Australia (ARC, 2014) consider citations for some subject areas, and there are advocates of increasing use of citations for some types of science when the results correlate because citation metrics are much cheaper than peer review (Abramo, Cicero, & D'Angelo, 2013; Franceschet & Costantini, 2011; Mryglod et al., 2013), although no simple method is likely to work (Butler & Mcallister, 2011). In addition, citations are used for formal and informal evaluations of academics (Cole, 2000) and the Journal Impact Factor (JIF) is a widely recognised and used citation metric.

The range of metrics of relevance to science has recently increased with the emergence of webometrics (Almind & Ingwersen, 1997), which includes a range of new impact indicators derived from the web (Kousha & Thelwall, 2014) and

* Corresponding author. Tel.: +44 1902 321470.
E-mail addresses: m.thelwall@wlv.ac.uk (M. Thelwall), PaulJWilson@wlv.ac.uk (P. Wilson).

http://dx.doi.org/10.1016/j.joi.2014.09.011
1751-1577/

altmetrics (Priem, Taraborelli, Groth, & Neylon, 2010), which incorporate many attention and impact indicators derived from social web sites (Priem, 2014). Altmetrics seem particularly promising to help researchers to identify recently published articles that have attracted a lot of attention (Adie & Roe, 2013) and to give evidence of non-standard impacts of research that can be added to CVs (ACUMEN, 2014; Piwowar & Priem, 2013). Statistical analyses of some of these have started to generate new insights into how science works (Mohammadi & Thelwall, 2014; Thelwall & Maflahi, in press) and the types of research impacts that are not recognised by traditional citation counts (Mohammadi & Thelwall, 2013).

Because of the many uses of citations within science, it is important to understand as much as possible about why they are created and why one article, researcher or research group may be more cited than another. Whilst citations may be given to acknowledge relevant previous work (Merton, 1973), they can also be used to criticise or argue against it (MacRoberts & MacRoberts, 1996) and so citations are not universally positive. Moreover, citations do not appear to be chosen in a dispassionate, unbiased way (Borgman & Furner, 2002). For example, researchers in some fields tend to cite papers written in the same language (Yitzhaki, 1998), highly relevant citations may be systematically ignored (McCain, 2012) and fame seems also to attract citations (Merton, 1968). There are also field differences in the typical number of citations received by papers (e.g., Waltman, van Eck, van Leeuwen, Visser, & van Raan, 2011) and review articles are more likely to be highly cited than other articles (e.g., Aksnes, 2003). From a different perspective, factors that associate with highly cited papers, such as collaboration, internationality and referencing patterns (Didegah & Thelwall, 2013b; Sooryamoorthy, 2009), are important because they may push researchers and funders towards more successful types of research.

Although some of the factors affecting citations discussed above have been discovered using qualitative methods, such as interviews with authors, statistical methods are needed to identify the magnitude of factors and to detect whether they apply in particular cases. The simplest approach is probably to compare the average number of citations (or citation-based indicators) for one set of papers against that of another to see which is higher (van Raan, 1998). Another strategy is to assess whether citation counts correlate significantly against another metric that is hypothesised to be related (Shema, Bar-Ilan, & Thelwall, 2014). The most discriminating methods used so far are regression-based because they allow the effects of multiple variables to be examined simultaneously. In particular, regression guards against one factor being identified as significant (e.g., international collaboration) when another factor (e.g., collaboration with the USA) is underlying cause of higher (or lower) citations.

There is no consensus about the best regression method for citation data. Methods used so far include ordinary least squares linear regression (Aksnes, Rørstad, Piro, & Sivertsen, 2013; Dragos & Dragos, 2014 [citations per publication used as the dependant variable]; Foo & Tan, 2014; He, 2009; Mavros et al., 2013; Rigby, 2013 [adding 1 to citations, dividing by a time normalised value and taking their log]; Tang, 2013 [adding 1 to citations and taking their log]; Stewart, 1983), logistic regression (Baldi, 1998; Bornmann & Williams, 2013; Kutlar, Kabasakal, & Ekici, 2013; Sin, 2011; Willis, Bahler, Neuberger, & Dahm, 2011; Xia & Nakanishi, 2012; Yu, Yu, & Wang, 2014), a distribution-free regression method (Peters & van Raan, 1994), multinomial logistic regression (Baumgartner & Leydesdorff, 2014) and negative binomial regression (Chen, 2012; Didegah & Thelwall, 2013a, 2013b; McDonald, 2007; Thelwall and Maflahi, in press [for altmetrics]; Walters, 2006; Yoshikane, 2013 [for patent citations]).

The typical distribution of citations is highly skewed (de Solla Price, 1965; Seglen, 1992), so that tests based upon the normal distribution (e.g., ordinary least squares regression) are not appropriate if the data is raw citation counts. Logistic regression can avoid this issue by predicting membership of the highly cited group of papers rather than directly predicting citations. Whilst negative binomial regression can cope with skewed data and is designed for discrete numbers (Hilbe, 2011), the most appropriate distribution for citations to a collection of articles from a single subject and year seems to be the discrete lognormal distribution (Evans, Hopkins, & Kaube, 2012; Radicchi, Fortunato, & Castellano, 2008; Thelwall & Wilson, 2014) and the hooked power law is also a reasonable choice (Thelwall & Wilson, 2014). Although many articles suggest a power law for the tail of citation distributions (e.g., Yao, Peng, Zhang, & Xu, 2014) this is not helpful for statistical analyses that need to include all cited articles and is broadly consistent with a lognormal distribution for all articles, although small discrepancies may be revealed by fine-grained analyses (Golosovsky & Solomon, 2012). Citations to articles from individual journals almost always conform to a lognormal distribution (Stringer, Sales-Pardo, & Amaral, 2010), as do some other citation-based indicators also follow a lognormal distribution (e.g., generalised *h*-indices: Wu, 2013). Although it has not been fully tested, it seems likely that most sets of articles from a specific time window will approximately follow a discrete lognormal distribution, unless the time window is too long or very recent. Hence it is not clear that negative binomial regression is optimal when the dependant variable is citation counts.

Neither the discrete lognormal or the hooked power law distributions have been used for regression because it seems that no software exists for this. An alternative strategy would be to take the log of the citations and then use the general linear model to analyse them with the assumption that the logged citations would be normally distributed (the general linear model assumes that the error terms or residuals are normally distributed). Although the log of the continuous version of the lognormal distribution is a perfect normal distribution, the same is not true for the discrete lognormal distribution and so it is not obvious that this will work. Moreover, the use of log transformation for citation data has been argued against for classification purposes because of the variance reduction that it introduces (Leydesdorff & Bensman, 2006), but this is not evidence that it will not work for regression. This article assesses both of these and the continuous normal distribution in order to identify the most powerful, regression-based approach for citations and similar data, such as altmetrics. The results will help to ensure that future statistical analyses of the factors affecting citation counts are as powerful and reliable as possible.