Journal of Informetrics 9 (2015) 499–513

Contents lists available at ScienceDirect

Journal of Informetrics

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

Modelling count response variables in informetric studies: Comparison among count, linear, and lognormal regression models

Isola Ajiferuke[a,*], Felix Famoye[b]

[a] Faculty of Information and Media Studies, University of Western Ontario, London, Canada N6A 5B7
[b] Department of Mathematics, Central Michigan University, Mount Pleasant, MI 48859, USA

ARTICLE INFO

Article history:
Received 7 January 2015
Received in revised form 10 May 2015
Accepted 11 May 2015

Keywords:
Count response variable
Linear regression model
Count regression models
Negative binomial regression model
Lognormal regression model
Informetric studies

ABSTRACT

The purpose of the study is to compare the performance of count regression models to those of linear and lognormal regression models in modelling count response variables in informetric studies. Identified count response variables in informetric studies include the number of authors, the number of references, the number of views, the number of downloads, and the number of citations received by an article. Also of a count nature are the number of links from and to a website. Data were collected from the United States Patent and Trademark Office (www.uspto.gov), an open access journal (www.informationr.net/ir/), Web of Science, and Maclean's magazine. The datasets were then used to compare the performance of linear and lognormal regression models with those of Poisson, negative binomial, and generalized Poisson regression models. It was found that due to over-dispersion in most response variables, the negative binomial regression model often seems to be more appropriate for informetric datasets than the Poisson and generalized Poisson regression models. Also, the regression analyses showed that linear regression model predicted some negative values for five of the nine response variables modelled, and for all the response variables, it performed worse than both the negative binomial and lognormal regression models when either Akaike's Information Criterion (AIC) or Bayesian Information Criterion (BIC) was used as the measure of goodness of fit statistics. The negative binomial regression model performed significantly better than the lognormal regression model for four of the response variables while the lognormal regression model performed significantly better than the negative binomial regression model for two of the response variables but there was no significant difference in the performance of the two models for the remaining three response variables.

© 2015 Elsevier Ltd. All rights reserved.

1. Introduction

In many studies, a linear regression model is often employed for parameter estimation, goodness of fit, or response variable prediction. In a linear regression model, the response variable follows a continuous normal distribution which implies that predicted values could be any real value that is negative or positive. However, for many experiments, survey, observations, etc., the response variables are of a count nature, i.e. they follow a discrete (or count) distribution. Using a

[*] Corresponding author. Tel.: +1 519 661 2111x81364; fax: +1 519 661 3506.
E-mail addresses: iajiferu@uwo.ca (I. Ajiferuke), felix.famoye@cmich.edu (F. Famoye).

http://dx.doi.org/10.1016/j.joi.2015.05.001
1751-1577/© 2015 Elsevier Ltd. All rights reserved.

linear regression model for such variables might lead to negative predicted values. So, how has this problem been addressed in the literature? The approach in many fields of study is to use a count regression model instead of a linear regression model. For example, in health-related studies, count regression models have been used to model the number of incidents of physical aggression or substance abuse (Gagnon, Doron-LaMarca, Bell, O'Farrell, & Taft, 2008), the number of malaria cases (Achcar, Martinez, Pires de Souza, Tachibana, & Flores, 2011), the number of medically attended childhood injuries (Karazsia & van Dulmen, 2008), number of health benefits received per patient (Czado, Schabenberger, & Erhardt, 2014), and number of sub-health symptoms (Xu, Li, & Chen, 2011). In other fields of study, they have also been used to estimate recreational trip demands (Wang, Li, Little, & Yang, 2009), number of auto insurance claims (Meng, 2009), number of roadway accidents (Nassiri, Najaf, & Amiri, 2014), and number of hardware failures or occurrences of disease or death (Gulkema & Goffelt, 2008).

In informetric studies, especially in the subfields of citation analysis, patent analysis, webometrics, and altmetrics, possible count response variables include: number of papers published by a scholar, institution, or country; number of papers collaborated by scholars, institutions, or countries; number of citations received by a paper; number of times two papers are co-cited; number of other patents referencing a patent; number of inlinks to a website; number of co-links to two websites; and number of views or downloads received by an online paper. So, how have informetric studies modelled these count response variables? Many studies were only interested in the correlation between these variables (Bornmann, Schier, Marx, & Daniel, 2012; Buter & van Raan, 2011; Guerrero-Bote & Moya-Anegón, 2014; Jamali & Nikzad, 2011; Kim, 1998; Kreider, 1999; Liu, Fang, & Wang, 2011; Moed, 2005; Nieder, Dalhaug, & Aandahl, 2013; Schlögl & Gorraiz, 2010; Schlögl, Gorraiz, Gumpenberger, Jack, & Kraker, 2014; Tsay, 1998; Yuan & Hua, 2011) and some other variables while a couple of studies have employed logistic regression models by recoding the count variable into a dichotomous variable (Gargouri et al., 2010; Willis, Bahler, Neuberger, & Dahm, 2011).

Many other studies have also employed the linear regression to model these count variables (Ajiferuke, 2005; Ayres & Vars, 2000; Bornmann & Daniel, 2007; Habibzadeh & Yadollahie, 2010; Landes & Posner, 2000; Lokker, McKibbon, McKinlay, Wilczynski, & Haynes, 2008; Peters & van Raan, 1994; Vaughan & Thelwall, 2005; Willis et al., 2011; Xia, Myers, & Wilhoite, 2011; Yoshikane, Suzuki, Arakama, Ikeuchi, & Tsuji, 2013; Yu, Yu, Li, & Wang, 2014). However, negative binomial regression models, which are a type of count regression models, have been used in a very few informetric studies (Baccini, Barabesi, Cioni, & Pisani, 2014; Chen, 2012; Didegah & Thelwall, 2013; Lee, Lee, Song, & Lee, 2007; McDonald, 2007; Thelwall & Maflahi, 2015; Walters, 2006; Yoshikane, 2013; Yu & Wu, 2014). These studies, except the one by Yoshikane, did not compare the performance of negative binomial regression models and linear regression models. In the case of Yoshikane, the negative binomial regression and logistic regression models were used to confirm the significant factors found in the linear regression model for patents' cited frequency. Also, in a recent article by Thelwall and Wilson (2014), the abilities of negative binomial regression, lognormal regression and general linear regression models in detecting factors affecting citation scores were compared. Assuming that citation counts tend to follow a discrete lognormal distribution, the authors simulated discrete lognormal citation data and regressed it with one binary factor. The results of the study showed that "negative binomial regression applied to discrete lognormal data will identify non-existent factors at a higher rate than expected by the significance level" (p. 969), and the authors recommended the following strategy for citation data that follows a discrete lognormal distribution: take the logarithm of the citation data after discarding zeros and then apply the general linear model OR add 1 to the data before taking the logarithm and then use the general linear model. It should be noted that: not all informetric response variables follow discrete lognormal distribution (in fact, not even all citation data follow the discrete lognormal distribution); the above study made use of simulated data, and not real data; and the study used only one simple factor, and no continuous covariates or multiple factors as is often the case in major studies. Furthermore, there are other reasons, apart from the ability to detect factors affecting the response variable, why one regression model may be preferred to another. Hence, the objectives of this paper are to:

- Illustrate the pitfalls of using linear regression models for count response variables with real data sets, especially given their frequent use in empirical studies by informetric researchers;
- Investigate the suitability of lognormal regression model in modelling response variables with multiple covariates/factors using real data sets; and
- Compare the performance of linear, count, and lognormal regression models in modelling informetric count response variables.

2. Linear, lognormal, and count regression models

A brief review of linear, lognormal, and three count regression models will be given in this section.

2.1. Linear regression

A linear regression model stipulates that the response variable y can be written as

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{p-1} x_{i,p-1} + \varepsilon_i, \quad i = 1, 2, 3, \ldots, n \tag{1}$$

where xs are the predictors, βs are the regression parameters and error ε_i is assumed to have a normal distribution with mean 0 and constant variance σ^2. Hence, the response variable y has mean $E(Y_i|x_i) = \mu(x_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{p-1} x_{i,p-1}$

Download English Version:

https://daneshyari.com/en/article/10358526

Download Persian Version:

https://daneshyari.com/article/10358526

Daneshyari.com