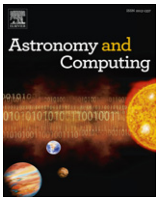
Astronomy and Computing 12 (2015) 21–32

Contents lists available at ScienceDirect

Astronomy and Computing

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



Full length article

The overlooked potential of Generalized Linear Models in astronomy, I: Binomial regression

R.S. de Souza[a,*], E. Cameron[b], M. Killedar[c], J. Hilbe[d,e], R. Vilalta[f], U. Maio[g,h], V. Biffi[i], B. Ciardi[j], J.D. Riggs[k], for the COIN collaboration

[a] *MTA Eötvös University, EIRSA "Lendulet" Astrophysics Research Group, Budapest 1117, Hungary*
[b] *Department of Zoology, University of Oxford, Tinbergen Building, South Parks Road, Oxford, OX1 3PS, United Kingdom*
[c] *Universitäts-Sternwarte München, Scheinerstrasse 1, D-81679, München, Germany*
[d] *Arizona State University, 873701, Tempe, AZ 85287-3701, USA*
[e] *Jet Propulsion Laboratory, 4800 Oak Grove Dr., Pasadena, CA 91109, USA*
[f] *Department of Computer Science, University of Houston, 4800 Calhoun Rd., Houston TX 77204-3010, USA*
[g] *INAF — Osservatorio Astronomico di Trieste, via G. Tiepolo 11, 34135 Trieste, Italy*
[h] *Leibniz Institute for Astrophysics, An der Sternwarte 16, 14482 Potsdam, Germany*
[i] *SISSA — Scuola Internazionale Superiore di Studi Avanzati, Via Bonomea 265, 34136 Trieste, Italy*
[j] *Max-Planck-Institut für Astrophysik, Karl-Schwarzschild-Str. 1, D-85748 Garching, Germany*
[k] *Northwestern University, Evanston, IL, 60208, USA*

ARTICLE INFO

Article history:
Received 26 September 2014
Accepted 2 April 2015
Available online 29 April 2015

Keywords:
Cosmology
First stars
Methods
Statistical
Stars
Population III

ABSTRACT

Revealing hidden patterns in astronomical data is often the path to fundamental scientific breakthroughs; meanwhile the complexity of scientific enquiry increases as more subtle relationships are sought. Contemporary data analysis problems often elude the capabilities of classical statistical techniques, suggesting the use of cutting edge statistical methods. In this light, astronomers have overlooked a whole family of statistical techniques for exploratory data analysis and robust regression, the so-called Generalized Linear Models (GLMs). In this paper – the first in a series aimed at illustrating the power of these methods in astronomical applications – we elucidate the potential of a particular class of GLMs for handling binary/binomial data, the so-called logit and probit regression techniques, from both a maximum likelihood and a Bayesian perspective. As a case in point, we present the use of these GLMs to explore the conditions of star formation activity and metal enrichment in primordial minihaloes from cosmological hydro-simulations including detailed chemistry, gas physics, and stellar feedback. We predict that for a dark mini-halo with metallicity $\approx 1.3 \times 10^{-4} Z_{\odot}$, an increase of 1.2×10^{-2} in the gas molecular fraction, increases the probability of star formation occurrence by a factor of 75%. Finally, we highlight the use of receiver operating characteristic curves as a diagnostic for binary classifiers, and ultimately we use these to demonstrate the competitive predictive performance of GLMs against the popular technique of artificial neural networks.

1. Introduction

The simple *linear regression* model has long been a mainstay of astronomical data analysis, the archetypal problem being to determine the line of best fit through Hubble's diagram (Hubble, 1929). In this approach, the expected value of the response variable, $\boldsymbol{Y} \in \mathbb{R}^m$, is supposed linearly dependent on its coefficients, $\boldsymbol{\beta} \in \mathbb{R}^n$, acting upon the set of n predictor variables, $\mathbf{X} \in \mathbb{R}^{n \times m}$,

$$E(\boldsymbol{Y}) = (\boldsymbol{\beta}^T \mathbf{X})^T. \qquad (1)$$

The least-squares fitting procedure for performing this type of regression (Isobe et al., 1990) relies on a number of distributional assumptions which fail to hold when the data to be modelled come from *exponential family* distributions other than the Normal/Gaussian (Hardin and Hilbe, 2012; Hilbe, 2014). For instance, if the response variable takes the form of Poisson distributed count data (e.g. photon counts from a CCD), then the equidispersion property

[*] Corresponding author.
E-mail addresses: rafael.2706@gmail.com (R.S. de Souza), dr.ewan.cameron@gmail.com (E. Cameron).

http://dx.doi.org/10.1016/j.ascom.2015.04.002
2213-1337/© 2015 Elsevier B.V. All rights reserved.

of the Poisson, which prescribes a local variance equal to its conditional mean, will directly violate the key linear regression assumption of *homoscedasticity* (a common global variance independent of the linear predictors). Moreover, adopting a simple linear regression in this context means to ignore another defining feature of the Poisson: its ability to model data with only non-negative integers. Similar concerns arise for modelling Bernoulli and binomial distributed data (i.e., on/off, yes/no) where regression methods optimized for continuous and unbounded response variables are of limited assistance (Hilbe, 2009).

Yet, data analysis challenges of this sort arise routinely in the course of astronomical research: for example, in efforts to characterize exoplanet multiplicity as a function of host multiplicity and orbital separation (Poisson distributed data; Wang et al., 2014), or to model the dependence of the galaxy bar fraction on total stellar mass and redshift (Bernoulli distributed data; Melvin et al., 2014). For such regression problems there is a powerful solution already widely-used in medical research (e.g., Lindsey, 1999), finance (e.g., de Jong and Heller, 2008), and healthcare (e.g., Griswold et al., 2004) settings, but vastly under-utilized to-date in astronomy. This is known as Generalized Linear Models (GLMs). Basic GLMs include Normal or Gaussian regression, gamma and inverse Gaussian models, and the discrete response binomial, Poisson and negative binomial models.

1.1. Generalized linear models

The class of GLMs, first developed by Nelder and Wedderburn (1972), take a more general form than in Eq. (1):

$$E(\mathbf{Y}) = g^{-1}\left((\boldsymbol{\beta}^T\mathbf{X})^T\right), \tag{2}$$

with the response variable, $\mathbf{Y} \mid \boldsymbol{\beta}^{\mathbf{T}}\mathbf{X}$, belonging to a specified distribution from the single parameter exponential family and $g^{-1}(\cdot)$ providing an appropriate transformation from the linear predictor, $(\boldsymbol{\beta}^T\mathbf{X})^T$, to the conditional mean, μ. The inverse of the *mean function*, $g^{-1}(\cdot)$, is known as the *link function*, $g(\cdot)$. Nelder and Wedderburn (1972) and McCullagh and Nelder (1989) laid the foundations of the GLM estimation algorithm, which is a subset of maximum likelihood estimation. The algorithm they devised in early software development is for the most part still used today in the majority of GLM implementations—both in commercial statistical packages (e.g. SPSS and SAS) and in freeware-type packages (e.g. R and Python).

GLMs have received a great deal of attention in the statistical literature. Variations and extensions of the traditional algorithm have resulted in methodologies, such as: generalized estimating equations (Liang and Zeger, 1986); generalized additive models (Hastie and Tibshirani, 1986); fixed and random effects regression (Breslow and Clayton, 1993); quasi-least squares regression (Shults and Hilbe, 2014); and more. Bayesian statisticians working within the GLM framework have explored Gibbs sampling techniques for posterior sampling (Albert and Chib, 1993), various issues of prior choice (Gelman et al., 2008) and prior-sensitivity analysis (Doss and Narasimhan, 1994), developed *errors-in-variables* treatments (for the case of errors in the predictor variables; e.g. Richardson and Gilks, 1993 and Mallick and Gelfand, 1996), and devised Gaussian process-based strategies for the use of GLMs in geospatial statistics (Diggle et al., 2002). The GLM methodology thus stands at the base of a wide number of contemporary statistical methods.

Despite the ubiquitous nature of GLMs in general statistical applications, there have been only a handful of astronomical studies applying GLM techniques such as logistic regression (e.g. Raichoor and Andreon, 2012, 2014 and Lansbury et al., 2014), Poisson regression (e.g. Andreon and Hurn, 2010); and the importance of modelling overdispersion in count data (as facilitated by the negative binomial GLM) has only lately become appreciated through cosmological research (Ata et al., 2015). Hence, in this series of papers we aim to demonstrate the vast potential of GLMs to assist with both exploratory and advanced astronomical data analyses through the application to a variety of astronomical inference problems.

The astronomical case studies explored herein focus on an investigation of the statistical properties of baryons inside simulated high-redshift haloes, including detailed chemistry, gas physics and stellar feedback. The response variables are categorical with two possible outcomes and therefore Bernoulli distributed. In our particular case, these correspond to either (i) the presence/absence of star formation activity, or (ii) metallicity above/below the critical metallicity ($Z_{\rm crit}$) associated with the first generation of stars. The predictor variables are properties of high-redshift galaxies with continuous domain.

The outline of this paper is as follows. In Section 2 we describe the cosmological simulation and the dataset of halo properties. We describe various forms of binomial GLM regression in Section 3. In Section 4 we present our analysis of the simulated dataset for the two selected response variables. In Section 5 we discuss critical diagnostics of our analysis, and compare our classifications with those that use artificial neural networks in Section 6. Finally, in Section 7 we summarize our conclusions.

2. Simulations

In order to ascertain the key ingredients that affect star formation in the early Universe, we study cosmological simulations of high-redshift galaxies and proto-galaxies. In the following, we describe the simulated data used to exemplify the unique benefits of binomial GLM regression for modelling galaxy properties that are naturally addressed as a dichotomous problem.

2.1. Runs

The data set used in this work is retrieved from a cosmological hydro-simulation based on Biffi and Maio (2013) (see also Maio et al., 2010, 2011 and de Souza et al., 2014). The code employed to run the simulation is GADGET-3, a modified version of the parallel N-body, smoothed-particle hydrodynamics code named GADGET-2 (Springel, 2005). The modifications include: a relevant chemical network to self-consistently follow the evolution of different atomic and molecular chemical species (e.g., Yoshida et al., 2003; Maio et al., 2006, 2007, 2009); metal pollution according to proper stellar yields and lifetimes for both the pristine population III (Pop III) and the following population II/I (Pop II/I) star forming regime (Tornatore et al., 2007; Maio et al., 2010); radiative gas cooling from molecular, resonant and fine-structure lines (Maio et al., 2007). The actual stellar population is determined by the local heavy-element mass fraction (metallicity, Z) and the existence of a critical threshold $Z_{\rm crit} = 10^{-4}Z_\odot$[1] (e.g., Omukai, 2000; Bromm et al., 2001) below which Pop III star formation takes place and above which Pop II/I stars are formed.

The initial matter density field is sampled at redshift $z = 100$ adopting the standard cold dark matter model with cosmological constant Λ, ΛCDM. The cosmological parameters at the present time are assumed to be: $\Omega_{0,\Lambda} = 0.7$, $\Omega_{0,m} = 0.3$, $\Omega_{0,b} = 0.04$, for cosmological-constant, matter and baryon density, respectively

[1] Despite the uncertainties on $Z_{\rm crit}$, it is safe to assume values around $Z_{\rm crit} = 10^{-4}Z_\odot$, in fact even order-of-magnitude deviations would not change significantly the final results in terms of star formation and cosmic metal pollution (see details in Maio et al., 2010).