The Overlooked Potential of Generalized Linear Models in Astronomy - I: Binomial Regression and Numerical Simulations

R S. de Souza\textsuperscript{a}, E. Cameron\textsuperscript{b}, M. Kille
dar\textsuperscript{c}, J. Hilbe\textsuperscript{d,e}, R. Vilalta\textsuperscript{f}, U. Maio\textsuperscript{g,h}, V. Biffl\textsuperscript{i}, B. Ciardi\textsuperscript{j}, J. D. Riggs\textsuperscript{k}, for the COIN collaboration

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

Abstract

Revealing hidden patterns in astronomical data is often the path to fundamental scientific breakthroughs; meanwhile the complexity of scientific inquiry 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. 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.

Keywords: cosmo
gical calculations – scientific computation – cosmology

1. Introduction

In the simple linear regression model, the expected value of the response variable, \( Y \) (an \( \mathbb{R}^m \) column vector), is supposed linearly dependent on its coefficients, \( \beta \) (an \( \mathbb{R}^m \) column vector), acting upon the set of \( n \) predictor variables, \( X \) (an \( n \times m \) matrix), i.e., \( E(Y) = (\beta^T X)^T \). This approach has long been a mainstay of astronomical data analysis recipes, the archetypal linear regression problem being to determine the line of best fit through Hubble’s diagram (Hubble, 1929). 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).
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 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 exists 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 the form:

$$E(Y) = g^{-1}((\beta^T X)^T),$$

with the response variable, $Y | \beta^T X$, belonging to a specified distribution from the single parameter exponential family and $g^{-1}()$ providing an appropriate transformation from the linear predictor, $(\beta^T X)^T$, to the conditional mean. The inverse of the mean function, $g^{-1}()$, is known as the link function, $g()$. 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).

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; 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. Lansbury et al. 2014) and 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., 2014). 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 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 \(Z_{\text{crit}}\) 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 §2 we describe the cosmological simulation and the dataset of halo properties. We describe various forms of binomial GLM regression in §3. In §4 we present our analysis of the simulated dataset for the two selected response variables. In §5 we discuss critical diagnostics of our analysis, and compare our classifications with those that use artificial neural networks in §6. Finally in §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 Bih and Maio 2013 (see also Maio et al. 2010, 2011, de Souza et al., 2014). The code employed to run the simulation is 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, such as \(e^-\), H, H\(^+\), H\(^++\), He, He\(^+\), He\(^{++}\), H\(_2\), H\(_2^+\), D, D\(^+\), HD, HeH\(^+\) (e.g., Yoshida et al., 2003; Maio et al., 2006, 2007, 2009); metal pollution according to proper stellar yields (He, C, O, Si, Fe, Mg, S, etc.) 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).

Stellar population properties are matter of debate, since masses, lifetimes and metal yields of primordial stars are still uncertain. Broadly speaking, one can identify a primordial (Pop III) regime by simply looking at the abundance of heavy elements (metallicity, \(Z\)) of the hosting star forming environment. Early stars are expected to originate in pristine or low-\(Z\) gas, while more standard stars (Pop II/I), similar to those observed in the local Universe, are preferentially formed in metal-enriched gas. Thus, we rely on the existence of a critical metallicity \(Z_{\text{crit}}\) (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. Despite the uncertainties on \(Z_{\text{crit}}\), it is safe to assume values around \(Z_{\text{crit}} = 10^{-4}Z_{\odot}\), 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). The primordial Pop III initial mass function (IMF) is chosen to be top-heavy over the stellar mass range \([100, 500] M_{\odot}\) with a power-law distribution and a Salpeter (1955) slope of \(-2.35\). Such IMF allows for enrichment of heavy elements by powerful explosions of massive pair-instability supernovae in the range \([140, 260] M_{\odot}\) (e.g. Heger and Woosley, 2002), possibly emitting gamma rays over cosmological epochs (Maio and Barkov, 2014). The Pop II/I IMF follows a Salpeter shape over the range \([0.1, 100] M_{\odot}\) and corresponding supernovae that are further responsible for cosmic metal spreading (mainly C, O and \(\alpha\) elements) have masses above roughly \(40 M_{\odot}\).

The initial matter density field is sampled at redshift \(z = 100\) adopting the standard cold dark matter model with cosmological constant \(\Lambda\), 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, respect The expansion parameter at the present day is assumed to be \(H_0 = 100 h \text{ km/s/Mpc},\) with \(h = 0.7,\) while the primordial power spectrum has a slope \(n = 1\) and is normalized by imposing a mass variance within the 8-kpc/h sphere radius of \(\sigma_8 = 0.9\).

We consider snapshots in the range \(9 \lesssim z \lesssim 19,\) for a cubic volume of comoving side \(\sim 0.7\) Mpc, sampled with \(2 \times 320^3\) particles per gas and dark-matter
species. The resulting resolution is \(42 \, M_{\odot} h^{-1}\) and \(275 \, M_{\odot} h^{-1}\) for gas and dark matter, respectively.

### 2.2. Data set

We identify simulated objects by applying a friends-of-friends (FoF) algorithm with a linking length equal to 20 per cent of the mean inter-particle separation, and substructures are identified by using a SubFind algorithm (e.g. [Dolag et al., 2009] and references therein) which discriminates among bound and non-bound particles. The halo characteristics, such as position, velocity, dark matter, baryonic and chemical properties are computed and stored at each redshift.

The simulation outputs considered here consist of six parameters: dark-matter mass \((M_{\text{dm}})\), gas mass \((M_{\text{gas}})\), stellar mass \((M_{\text{star}})\), star formation rate \((SFR)\), metallicity \((Z)\) and gas molecular fraction \((x_{\text{mol}})\).

The total data set encompasses a few thousand primordial haloes at very high redshift, \(z \approx 19\), and comprises \(\sim25000\) objects at \(z \approx 9\). Out of these, we restrict our selection to those objects in which the gas component is resolved with at least 300 particles, so that convergent results are guaranteed and numerical artefacts due to poor statistics are avoided ([Bate and Burkert, 1997]). This typically corresponds to objects with at least \(\sim10^3\) particles in total. The resulting sample studied in this work is ultimately composed of 1680 haloes in the whole redshift range, with about 200 objects at \(z = 9\). The masses of the haloes are in the range \(10^3 M_{\odot} \lesssim M_{\text{dm}} \lesssim 10^5 M_{\odot}\), with corresponding gas masses between \(10^3 - 10^4 M_{\odot}\). The temperatures of these haloes typically range from 500 to \(10^4\) K and their thermal properties at such early times are mainly shaped by \(H_2\). Thermal effects due to supernova explosions are instead responsible for enriching and heating the gas in nearby smaller haloes, that therefore reach hotter temperatures \((\gtrsim 10^5\) K).

The interested reader can find in [Biffi and Maiolino, 2013] a more detailed discussion of the thermal and dynamical properties of the primordial objects analysed in this paper.

### 3. GLM Regression for Binary Response Data

In preparation for the application of binomial GLM regression we begin with a discussion of the two most common link functions: logit and probit (§3.1). Then we describe three variations on a class of GLMs which apply to binary response data: the maximum likelihood (ML) approach with logit link function (§3.2); and the Bayesian approach with a logit link function (§3.3) and with a probit link function amenable to exact Gibbs sampling (§3.4). These will be applied in the following section (§4) in the context of two specifically chosen astrophysical problems: i) presence/absence of star formation activity; ii) gas metallicity below/above \(Z_{\text{crit}}\) to discriminate between Pop III/Pop II/I star formation mode. The interested reader can find a comprehensive description of the underlying theory behind GLMs in [Zuur et al., 2013].

#### 3.1. Logit and probit regression

The Bernoulli distribution describes a process in which there are only two possible outcomes: success or failure (yes/no, on/off, red/blue, etc.; typically coded as 1/0)--the former occurring with probability, \(p\), and the other with probability, \(1 - p\). For multiple independent Bernoulli observations the total success count, \(k\), follows a binomial distribution:

\[
P(k) = \binom{n}{k} p^k (1 - p)^{n-k}.
\]

Both distributions are members of the exponential family (supposing the number of binomial trials, \(n\), is known and fixed) and thus may be used (equivalently) as the response distribution for modelling binary response data in the GLM framework.

The link function chosen in this case is designed to ensure a bijection\(^1\) between the \((-\infty, \infty)\) range of the linear predictor, \((\beta^T X)\), and the \((0,1)\) range of non-trivial probabilities for the binomial population proportion (the Bernoulli \(p\)). To this end there are two popular choices: the logit function:

\[
g(p) = \log\frac{p}{1-p},
\]

\[
\begin{align*}
\mu^T &= g^{-1}(\beta^T X) = \frac{\exp(\beta^T X)}{1 + \exp(\beta^T X)},
\end{align*}
\]  

\(^1\) See [Cameron, 2011] for a review of the binomial distribution and both maximum likelihood and Bayesian approaches to estimation of confidence/credible intervals on \(p\).

\(^2\) A function \(f\) from a set \(X\) to a set \(Y\) with the property that, for every \(y\) in \(Y\), there is exactly one \(x\) in \(X\) such that \(f(x) = y\).
and the probit function:
\[
\begin{align*}
g(p) &= \Phi^{-1}(p), \\
\mu^T &= g^{-1}(\beta^T X) = \Phi(\beta^T X),
\end{align*}
\]
where \( \Phi(\cdot) \) represents the Normal distribution function; the choice of link function defining the GLM as logit regression or probit regression, accordingly. Both link functions describe sigmoid curves smoothly and monotonically increasing from \( \mu = 0 \) at \( \beta^T X = -\infty \) to \( \mu = 1 \) at \( \beta^T X = \infty \) with the greatest rate of change occurring at \( \beta^T X = 0 \).

The logit function is most commonly preferred in clinical research applications where outcomes are most naturally described in terms of the odds-ratio, \( \frac{p}{1-p} \) (e.g. the relationship between the odds-ratio of patient recovery/non-recovery and the concentration of an administered drug); whereas the probit function is often presented within Bayesian statistical applications exploiting an associated Gibbs sampling algorithm. Sigmoid curves such as those described by the logit and probit functions may already be seen in empirical/phenomenological astronomical models: for example, in describing the fraction of quenched galaxies as a function of mass and/or environmental density (Peng et al., 2010; Rodriguez-Puebla et al., 2014).

A reason for employing logit or probit regression to model binary response data is to obtain for objects with only \( X \) observations, but no observed \( Y \)'s, the predicted probabilities that the unobserved response variable has the value 1 indicating “success”, however that is defined (e.g., “galaxy is quenched”, “star hosts planet”). Both models usually produce similar probabilities; though probit regression is not as commonly used for assessing the relationship of a predictor to the response since the interpretation of the exponentiated coefficient of a logit predictor as an odds ratio is a desirable feature of that model. Probit regression is normally used when a continuous variable to the response since the interpretation of the exponentiated coefficient of a logit predictor as an odds ratio is a desirable feature of that model. Probit regression is normally used when a continuous variable is dichotomized so that it becomes a binary response (see Zuur et al., 2013 for an examination of these and related issues with logit and probit models from both a frequency and Bayesian perspective).

3.2. Maximum-likelihood GLM regression with logit link function

Despite the growing popularity of Bayesian statistical analysis in the physical sciences the ML approach to GLM fitting remains the default in the majority of statistical software packages[3] for this reason, and its historical significance (cf. the extensive treatment given by McCullagh and Nelder 1989), we describe this approach first.

With the likelihood of the dataset fully specified by the linear predictor, \( \beta^T X \), and the choice of response variable distribution and link function of the GLM, the corresponding likelihood function for regression is both readily tractable and easily evaluated computationally. Iterative algorithms operating on the negative log-likelihood, such as the iteratively re-weighted least squares procedure used by glm (Venables and Ripley, 2002), thus provide a fast computational strategy for recovering the ML solution. The output from a standard ML GLM fitting code will typically be a list containing: (i) a ML estimate, \( \hat{\beta}_i \), for the \( \beta_i \) component of each candidate predictor variable, \( X_i \); (ii) the associated estimate of its standard error, \( \hat{\sigma}_{\beta_i} \), from which approximate confidence intervals on \( \beta_i \) may be obtained using the Normal distribution function (e.g. a 95% CI: \( \hat{\beta}_i \pm 2\hat{\sigma}_{\beta_i} \)); and (iii) a \( p \)-value computed from the Wald test using (i) and (ii), required for significance testing of the given predictor variable. Estimation of (ii) is by way of the observed information matrix according to asymptotic convergence theory for ML estimation.

In R the glm procedure may be called to perform ML estimation of the logistic regression model using the following general syntax:

```r
glm.fit <- glm(y~x1+x2+..., family = binomial("logit"))
```

The summary command can be called on the glm.fit object returned, as can plot which will display a number of useful fit and model checking diagnostics.

3.3. Bayesian GLM regression with logit link function

The bayesglm function in the CRAN[4] arm package is commonly used to estimate Bayesian logistic models (Gelman and Su, 2014). The code used to estimate this class of models is based on R’s default glm function. The basic syntax can be given as:

```r
bayesglm(y~x1+x2+..., family = binomial("logit"))
```

[3] Such as the glm in R
[4] http://cran.r-project.org
library(arm)
#Output identical to ML logit
blr1 <- bayesglm(y ~ x1+x2+ ..., family=binomial(link="logit"),
prior.scale=Inf, prior.df=Inf,
data=<datafile>)
display(blr1)

#Bayes GLM with default binomial
#logit link and Cauchy prior
#with scale=2.5
blr2 <- bayesglm(y~x1+x2+ ..., family=binomial,
data=<datafile>)
display(blr2)

#Bayes logit with normal prior
#with scale=2.5
blr3 <- bayesglm(y~x1+x2+ ..., family=binomial,
prior.scale=2.5, prior.df=Inf,
data=<datafile>)
display(blr3).

3.4. Bayesian GLM regression with probit link function

Use of the probit link function for Bayesian GLM regression has become a popular choice owing to
the availability of an exact Gibbs sampling algorithm for this model presented by [Albert and Chib(1993)].
The novelty of their algorithm is a data augmentation scheme in which an additional latent variable is added
for each observation having standard Normal distribution with mean set by the linear predictor, from which
the likelihood of the observed response is determined according to whether or not this latent variable is
above or below zero. Although the general sampler of the arm package does not in fact implement the
Albert and Chib (1993) scheme, it is important to note its availability for use in more complex Bayesian
hierarchial models built on the GLM framework (e.g. for the case of errors-in-variable GLM regression with
binary distributions for both predictor and response variable, such as can arise in comparing the sensitivity
two alternative tests). The basic syntax for using Bayesian probit GLM in R is given by:

library(arm)
bpr <- bayesglm(y~x1+x2+ ..., family=binomial(link="probit"),
prior.scale=2.5, prior.df=Inf,
data=<datafile>)
display(bpr).

The same criteria for the use of priors that we discussed for logit models above also maintain for
probit models. However, if the analyst desires to interpret the coefficients in terms of odds or risk ratios, a
logit model must be used, regardless if the model is based on ML or Bayesian methods. Predicted values
are nearly the same for logit or probit models.

4. Application to cosmological simulations

Within this section we demonstrate the application of the binomial regression techniques introduced
above to answer questions from an exploratory analysis of our cosmological hydro-simulation dataset that

---

5E.g., transformed to zero sample mean and unit sample variance as \( x_{new} = (x_{old} - \text{mean}(x_{old}))/\text{sd}(x_{old}) \). Where mean and sd represent the mean and standard deviation of the sample respectively.

6An alternative R function implementing the data augmentation scheme for Bayesian probit regression is available in the CRAN LearnBayes package (Albert [2007]) as bayes.probit.
could not be addressed by standard regression methods. Rather than exhaust all possible techniques for a single dataset, our aim is to demonstrate practical differences between distinct types of binary regression: i) Bayesian vs ML approach, both with the standard logit link (§3.2, 3.3); ii) Bayesian regression comparing logit vs probit link functions (§3.3, 3.4).

In the first case we consider the star formation activity connection with a preselected (physically motivated) set of predictor variables: \( x_{\text{mol}} \) and \( Z \). Alternatively, in our later analysis of the metallicity content of the galaxies, we use an automatic criterion to select the best choice of predictor variables among the entire set of halo features, or in other words the variable combination that minimizes the Akaike Information Criterion (AIC; Akaike, 1974). For all the following GLM analyses we quote the maximum likelihood (\( \mathcal{L}_{\text{max}} \)), the AIC as well as the alternative Bayesian Information Criterion (BIC; Schwarz, 1978).

4.1. Star formation activity

Here we discuss the connection between star formation activity and the gaseous chemical properties, \( x_{\text{mol}} \) and \( Z \) of proto-galaxies, using a Bayesian and a ML approach with logit link. The formation of the first metal-free stars in the Universe ended the cosmic dark ages (de Souza et al., 2011, 2012; Bromm, 2013; de Souza et al., 2013b; Whalen et al., 2013a,b) and began the production of elements heavier than lithium (Maio et al., 2010, 2013; Wise et al., 2014). Thus, a key problem in physical cosmology is to understand the environmental properties of such objects (e.g., de Souza et al., 2013a; Biffi and Maio, 2013; Salvaterra et al., 2013), born out of the pristine conditions leftover by the Big Bang.

As a visual exploration, Figure 1 shows the scatter of \( x_{\text{mol}} \) and \( Z \) coloured according to the presence of star formation activity. The objects located in the top left with high \( Z \) and very low \( x_{\text{mol}} \) are strongly displaced from the general trend, highlighting the effects of metal enrichment of quiescent galaxies polluted by external sources. The bottom right corner is not populated because gas with large molecular fractions of \( x_{\text{mol}} \sim 10^{-2} \) or higher would have very short cooling times, hence would immediately form stars which pollute the surrounding medium. Therefore, the larger the deviation from the general trend, the higher the effects of feedback mechanisms.

Figure 2 represents the distribution of \( x_{\text{mol}} \) and \( Z \) colour-coded by star formation activity and displayed by a box plot. The notches represent a rough guide of the uncertainty around the median of each distribution, \( \pm 1.58 \times \text{IQR}/\sqrt{n_{\text{obj}}} \), with \( n_{\text{obj}} \) being the number of objects, and IQR standing for interquartile range. A visual inspection suggests that \( x_{\text{mol}} \) plays a major role in triggering the star formation activity, in contrast to the lower influence of \( Z \) (see e.g., de Souza et al., 2014). The medians of haloes with and without star formation are different for both \( x_{\text{mol}} \) and \( Z \), indicating they might represent different populations, which reinforce their choice as predictor variables for star formation activity.

To perform the GLM analysis, we categorize the haloes via the binary response variable \( SFR_{\text{bin}} \), as those with \( (SFR > 0) \) and without \( (SFR = 0) \) star formation activity, a binary classification which makes it suitable for a binomial GLM analysis,

\[
SFR_{\text{bin}} = \begin{cases} 
1 \text{ or 'SF'} & \text{if } SFR > 0, \\
0 \text{ or 'no SF'} & \text{if } SFR = 0.
\end{cases}
\]
The underlying properties that act as predictor variables are: $x_{\text{mol}}$ and $Z$. We indicate with $p$ the probability that star formation activity is occurring in a galaxy. More specifically, $p = 1$ (0) if a galaxy has (has no) star formation. The predicted probability $\pi$ is then determined by the GLM analysis and compared to observed probability $p$ (for a given decision boundary), in order to ascertain the method’s performance, as explained below.

Table 1 shows the estimated coefficients and related $p$-values for the various linear predictors for both Bayesian and ML approach with the standard logit link. Since the variables have different units, we scaled the predictors by their means and divided by their standard deviations before the GLM analysis. As stated in §3, the GML analysis provides an estimate $\hat{\beta}_i$ for the $\beta_i$ component of each predictor variable. The values obtained can be used to calculate the fitted link function, $\eta$:

$$\eta = \hat{\beta}_0 + \hat{\beta}_1 Z + \hat{\beta}_2 x_{\text{mol}},$$

and transformed into a predicted probability, $\pi$:

$$\pi = \frac{e^\eta}{1 + e^\eta},$$

which can be used to assign a class membership for each object for a given probability decision threshold, $\pi_\text{th}$, i.e. SF = 1 if $\pi > \pi_\text{th}$ and 0 if $\pi < \pi_\text{th}$. For each halo, the predicted probability can be compared to the observed probability, which in this case is $p = 1$ if the halo presents star formation activity, and $p = 0$ otherwise. The performance of the method in reproducing the correct observed probabilities can be evaluated as
detailed in §5. When class sample sizes are approximately equal, which in this scenario would imply a similar number of galaxies with and without star formation activity, the optimal decision threshold is \( \pi_{th} \sim 0.50 \) (see §5.2).

Nevertheless, this criterion is not appropriate when the class sizes are imbalanced and an adjusting decision threshold has to be used. As a trivial example, if the data is imbalanced, the fit can predict \( \pi = 0.2 \) for all haloes with SF = 1 and \( \pi = 0.1 \) for all haloes with SF = 0. In this hypothetical scenario, the decision boundary would be in the range 0.1-0.2, instead of being 0.5 (50%) as one would naively expect. A more detailed explanation of how to adjust the decision threshold probability, \( \pi_{th} \), and a discussion of the predictive power of the method is given in §5.

The ML and Bayesian approaches give almost identical results for the estimated coefficients \( \hat{\beta} \), despite the addition of the prior. It seems that there is no preferred model, as indicated also by the comparison between the corresponding AIC, BIC and the logarithm of maximum likelihood \( L_{\text{max}} \). We note though the smaller credible intervals from the Bayesian logit in comparison to those from the ML analysis.

4.2. The Pop III-Pop II/I dichotomy

As previously mentioned, the first generation of stars (Pop III) are thought to form within pristine gas, while standard Pop II/I star formation takes place within metal enriched gas. Here we investigate the Pop III-PopII/I dichotomy using a Bayesian regression with logit and probit link functions.

Figure 3 shows the gas fraction versus molecular fraction with a colour scheme corresponding to stellar mass. A visual inspection indicates that larger molecular fractions are strongly associated with high-metallicity environments, confirming that the molecular fraction is the main predictor. From a physical point of view, the fact that the gas fraction in the environment of Pop II/I stars is usually lower than that of Pop III stars suggests that the cosmological production of early heavy elements enhances significantly gas cooling capabilities and boosts molecule formation in polluted material well above \( x_{\text{mol}} \sim \) a few percents. Basically, metal cooling allows gas fragmentation at regimes where pristine material is not able to condense – see the region: \( \{ x_{\text{mol}} > 10^{-2}, f_{\text{gas}} < 10^{-1.1} \} \).

The present cosmological simulations switch the stellar IMF from top-heavy to standard Salpeter when the metallicity exceeds \( Z_{\text{crit}} = 10^{-4}Z_\odot \) (see §2). To perform the GLM analysis in this section, we define \( Z_{\text{bin}} \) as the binary response variable, depending on whether the gas metallicity lies above or below \( Z_{\text{crit}} \):

\[
Z_{\text{bin}} = \begin{cases} 
1 & \text{or 'Pop II/I' if } Z \geq Z_{\text{crit}}, \\
0 & \text{or 'Pop III' if } Z < Z_{\text{crit}}.
\end{cases}
\] (7)

One can then use the binomial GLM regression to determine which global galaxy properties are linked to the dichotomy between the Pop II/I and Pop III host environment and how. We also use this problem as an opportunity to demonstrate the use of both logit \( \eta = \log \pi/(1-\pi) \) and probit \( \eta = \Phi^{-1}(\pi) \) link functions. Likewise the previous section, \( \pi \) here represents the predicted probability for the success of the binary response variable, in other words if a galaxy halo is an enriched Pop II/I environment given the underlying galaxy properties.

Firstly, one must identify the key galaxy properties as predictor variables. As in the previous example, we scale the predictors by their respective means and divide by the standard deviations before performing the analysis. Nonetheless, rather than adding a set of pre-chosen predictors, we herein illustrate a general feature selection approach, making use of \texttt{step} function in R. The method attempts to alternately drop and add members of an input set of predictor candidates in order to minimize the AIC of the fitted model. By using the stepwise algorithm we are able to select the most parsimonious combination of parameters and interaction terms from our input set\(^8\). In addition to the data-set described in §2.2, the predictor candidate parameters include: gas fraction, \( f_{\text{gas}} \equiv M_{\text{gas}}/M_{\text{dm}} \), stellar fraction, \( f_{\text{star}} \equiv M_{\text{star}}/M_{\text{dm}} \), and stellar-to-gas mass ratio \( M_{\text{star}}/M_{\text{gas}} \).

We found that \( x_{\text{mol}} \) plays the most important role in the predictive power of the model. Furthermore, the factors that maximise the information gain and are worth including as predictor parameters are: \( x_{\text{mol}}, f_{\text{gas}}, M_{\text{star}}/M_{\text{gas}}, \) and \( M_{\text{star}} \). The selection is equivalent

---

\(^8\)See also the \texttt{drop1} function in R, which is based on the likelihood ratio test.
regardless of whether the logit or probit link functions are used.

Having chosen suitable input variables, we can then apply the GLM analysis as described in §3.3 and §3.4. In Table 4.2 we provide the estimated coefficients for the predictor variables and respective \( p \)-values. The coefficients in the two cases are different as can be seen in Table 4.2, which is mostly a consequence of the different choices of link function. The predicted probabilities \( \pi \) are therefore estimated by solving the following equation:

\[
\eta = \hat{\beta}_0 + \hat{\beta}_1 x_{\text{mol}} + \hat{\beta}_2 f_{\text{gas}} + \hat{\beta}_3 M_{\text{star}} + \hat{\beta}_4 \frac{M_{\text{star}}}{M_{\text{gas}}},
\]

as well as either

\[
\Phi^{-1}(\pi) = \eta, \tag{9}
\]

if the probit link function is used, or

\[
\pi = \frac{e^\eta}{1 + e^\eta}, \tag{10}
\]

for the logit link function.

Ultimately, logit and probit regression result in similar predictions for the probability that the response variable is unity, i.e. \( \pi_{\text{logit}} \approx \pi_{\text{probit}} \). To illustrate this point, we calculate \( \pi \) twice for each galaxy in our sample given its underlying properties: once using the logit link and then again using the probit link. A histogram of the differences is shown in Figure 4. The logit link function leads to a value of \( \pi \) that is only slightly higher. Thus, for the case studied here, both link functions generate similar predictions, in spite of their different interpretations (see e.g., Zuur et al., 2013). A quantitative comparison between the predictive power of logit and probit, and the increase in number of relevant predictor variables are given in the following section.
5. Diagnostics

We now describe our experimental setting to assess the performance of GLM on the prediction of star formation activity $SFR_{\text{bin}}$ and metal enrichment $Z_{\text{bin}}$. We report on accuracy (i.e., fraction of events correctly classified) using a resampling technique known as 10-fold cross validation (Hastie et al., 2009) in §5.1, on Receiver Operating Characteristic (ROC) curves (Duda et al., 2000) in §5.2, and on the confusion matrix in §5.3.

5.1. Cross validation

When assessing model performance, it is of utmost importance to set aside a validation set to estimate the true generalization power of the model under analysis. This is particularly relevant to avoid the risk of model over-fitting. An over-fitted model captures aberrations on the training set that render the model useless during prediction. A popular approach to model validation makes use of resampling techniques (Hastie et al., 2009).

In the resampling technique known as $k$-fold cross validation, the data is divided into $k$ folds (subsamples) of equal size. The technique runs iteratively as follows. On each iteration, $k - 1$ folds are used for training (model fitting), while the remaining fold is used for testing (model assessment). The procedure repeats $k$ times, using mutually exclusive testing folds across iterations. The final result is the average over the score obtained on each iteration. Cross validation estimates the true performance of a classifier by exploiting all available information. In our experiments, we use a value of $k = 10$ to achieve a trade-off between bias (proportional to $k$) and variance (inversely proportional to $k$). Hereafter, all ROC curves and confusion matrices are estimated using the $k = 10$ cross-validation approach.

5.2. ROC curves

ROC curves provide both a visually and quantitative approach to report on the accuracy of predictions for binary classifiers. Hereafter, we refer to the classifications as positive (1) or negative (0). The technique consists of plotting the true positive rate (TPR or Sensitivity) vs the false positive rate (FPR or Specificity) as we vary the decision boundary $\pi_{\text{th}}$. The variation in the decision boundary enables us to assess the performance of the classifier under unequal error costs (i.e., under scenarios where the cost of a false positive is different from a false negative).

Specifically, to generate a ROC curve we make use of two measurements:

\[
\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}; \\
\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}},
\]

(11)

where TP = true positives, FP = false positives, TN = true negatives, and FN = false negatives. For example, in the case studied in §4.1, we would have:

- TP: the galaxy has SF and the method predicts SF,
- FP: the galaxy does not have SF but the method predicts SF,
- TN: the galaxy does not have SF and the method predicts no SF,
- FN: the galaxy has SF, but the method predicts no SF.
In this case the Sensitivity (Specificity) would quantify the ability of the method to correctly identify galaxies with (without) SF: the closer to 1 these values are, the more successful the analysis is. The same interpretation holds for the case discussed in §4.2 by replacing $SFR_{\text{bin}}$ with $Z_{\text{bin}}$.

Sensitivity is normally plotted on the $y$-axis, while $1 - \text{Specificity}$ is plotted on the $x$-axis (Figs. 5 and 6). The classifier is run several times with a different value of the decision threshold; each run provides a point in the (1-Specificity, Sensitivity) plane. The corresponding true ROC curve is obtained by joining the set of coordinates starting at (0,0) and ending at (1,1). An ideal ROC curve goes from (0,0) to (0,1) to (1,1). A quantitative approach to assess the quality of a ROC curve is to calculate the area under the curve (AUC), as a fraction of the area under the ideal curve, as often done in cases of discrepancy or inequality measurements (since e.g. Gini, 1912, 1921). Higher values of AUC correspond to more accurate classifiers, while a value of 0.5 corresponds to a random classifier (Hilbe, 2009).

The ROC curve can be used to access the optimal $\pi_{th}$, which is a trade-off between Sensitivity and Specificity. In order words, it is the one corresponding to the coordinate with minimum distance from (0,1), where both Sensitivity and Specificity are maximum. This is essential to ultimately assign a class membership for each data. A visual analysis of this classification scheme is made via confusion matrix, which will be discussed in the next section.

5.3. Confusion Matrix

A complementary diagnostics method is the confusion matrix $C$, which captures information about the actual and predicted classifications of a particular learning algorithm or classifier (Kohavi and Provost 1998). Columns in $C$ correspond to actual classes, whereas rows correspond to predicted classes (e.g., $SFR_{\text{bin}}, Z_{\text{bin}}$). The diagonal elements of the matrix contain the number of cases where the actual and predicted class agree, e.g., $C(i,i)$ contains the number of cases where class $i$ was predicted correctly. Off-diagonal elements capture all combination of misclassifications, e.g., $C(i,j)$ with $i \neq j$ contains the number of cases where class $i$ was incorrectly predicted as class $j$. On a $2 \times 2$ confusion matrix, entries along the diagonal stand for the number of true negatives TN (top left) and true positives TP (bottom right). Specifically, $C$ can be represented as follows:

$$
\begin{array}{c|cc}
& \text{TN} & \text{FN} \\
\hline
\text{FP} & \text{TP} \\
\end{array}
$$

6. Performance Comparison

During this section we compare the predictive performance of both logit vs probit links as discussed in §4.2 and between GLMs and artificial networks for the case discussed at §4.1.

6.1. Logit vs Probit

The left panel of Figure 5 shows a comparison between logit and probit ROC curves, pointing to the equivalence in predictive power of both methods, achieving an outstanding performance of $\text{AUC} = 0.95$, although their coefficients have a different interpretation.

In order to assess the relevance of a good set of predictor variables, the right panel of Figure 5 shows a visualization of the logit GLM regression obtained adding different predictor variables. While $f_{\text{gas}}, M_{\text{star}},$ and $M_{\text{star}}/M_{\text{gas}}$ together have a non-negligible contribution to explain the metallicity enrichment above/below $Z_{\text{crit}}$ with an $\text{AUC} = 0.74$, the molecular fraction, $x_{\text{mol}}$, clearly stands out as the most important parameter. This suggests that the level of molecular gas fraction has a strong connection with the level of metal content in primordial haloes.

6.2. Comparison between GLM and Neural Networks

We compare GLM with a popular non-parametric technique to classification known as Artificial Neural Networks (ANN) (Duda et al., 2000). A nonlinear multi-layer ANN is capable of expressing flexible decision boundaries over the variable space; it is a nonlinear statistical model that applies to both regression and classification. In particular, for an ANN with one hidden layer, each intermediate and

---

Note that as the ML and Bayesian approaches gives almost identical fitted coefficients, they lead to exactly same predicted probabilities.
output node computes a weighted combination of inputs, compressed (squashed) by a sigmoid (nonlinear) function (Bishop, 1996).

Figure 6 shows ROC curves for GLM and ANN analysis of the case presented in §4.1. The ROC curves were generated as those discussed in the previous section. In our experiments, GLM attains an AUC slightly higher than that of ANN (0.87 versus 0.83), reinforcing our claim for the competitiveness of GLM despite its inherent simplicity.

Figure 7 shows two confusion matrices, one for GLM (left) and one for ANN (right) for the case underlined in §4.1, i.e. the connection between star formation activity and the gaseous chemical properties \(x_{\text{mol}}\) and \(Z\) of galaxies. While TN is similar under both classifiers, TP differs significantly: ANN exhibits a high number of false negative FN (upper right) in contrast to the corresponding entry for GLM. Hence, the overall accuracy\(^{10}\) of GLM is 96.7%, while that of ANN is 93%.

7. Conclusions

We perform a comprehensive introduction of logit and probit generalized linear model regression for the astronomical community from both a maximum likelihood and a Bayesian perspective. As a real application, we analyse the host environment of the first generation of stars as predicted by numerical hydro-simulations of the early Universe, including detailed chemistry, gas physics, star formation, stellar evolution and stellar feedback. A summarizing flowchart visualization of the entire process is given in Appendix A.

The halo properties analysed here are categorical with two possible outcomes and therefore ideal candidates for the application of binomial GLM regression. These correspond to either (i) the presence/absence of star formation activity, or (ii) metal content above/below the critical metallicity associated to stellar population transition in primordial epochs.

In the first case, the explanatory variables were decided beforehand with preliminary physical motivation, while in the second case, we demonstrated

---

\(^{10}\) The accuracy is given by \(\frac{TN+TP}{TN+TP+FP+FN}\).
Figure 7: Confusion matrix for the model $\text{SFR}_{\text{bin}} \sim x_{\text{mol}} + Z$ (in R notation) discussed in §4.1 as expected by logistic GLM (left) and ANN (right). In each panel, the first and second columns refer to the simulated objects with (1532 galaxies) and without (148 galaxies) star formation activity, respectively. The on-diagonal elements refer to TN (top left) and TP (bottom right), while the off-diagonal elements refer to FP (top right) and FN (bottom left). The color scheme ranges from blue, correct values, to orange, incorrect values, with the intensity determined by the number of objects in each category.

the use of the AIC to select the most parsimonious set of variables from among a given set of candidates. This method is particularly beneficial for providing new insight into fundamental underlying galaxy properties.

A maximum likelihood as well as a Bayesian (with Cauchy priors) analysis result in very similar coefficients for each variable. We have explored the use of both logit and probit link functions and found that they lead to different $\hat{\beta}$ coefficients, but with the same sign. Never the less, calculations of the predicted probabilities produce very similar results regardless of whether a logit or probit model is used for estimation.

The GLM method has been shown to be very competitive against artificial neural networks, attaining an area under the curve (AUC) coefficient of 0.87 against 0.83 from ANN. Since a value of AUC = 1 indicates a perfect classifier and a value of AUC = 0.5 suggests a random predictor, both GLM and ANN approaches can be considered rather robust, albeit the AUC seems to favour slightly the GLM. Furthermore, given its inherently simplicity, GLM results are easily portable and have a more straightforward interpretation.

Also worth noting is that the potential of GLM regression goes far beyond binary classification. Many data situations involve discrete data, but are nevertheless modelled as if the response variable were continuous. If the data are modelled as discrete, it is by employing a Poisson model, without due regard for the corresponding distributional assumption of equality between mean and variance (equidispersion). This is a strongly restrictive technical assumption and is rarely met in real data. In practice, there are nearly as many count models as there are shapes of counts: there is a variety of mixture models, of zero-inflation models, of two-part hurdle models, of finite mixture models, etc. which assume that the counts being modelled are being generated from more than one source. Furthermore, there are non-parametric quantile count models, generalized additive models, models with endogenous stratification, 3-parameter count models and panel models, to name only a few. GLMs are of common use in the statistical literature, but almost Terra incognita in astronomical data analysis,
with only few recent notable applications of logistic regression (e.g. [Lansbury et al. 2014]), Poisson regression (e.g. [Arendon and Hurn 2010]) and negative binomial regression ([Ata et al. 2014]).

Finally, we highlight the vast potential of GLMs and extended GLMs for the astronomical community through their possible application to a plethora of astronomical problems, such as: photometric redshift estimation (gamma distributed data), globular cluster counts (Poisson distributed data), or galaxy morphological classification (multinomial distributed data). GLMs might be a precious instrument for astronomical investigations, thanks to their capabilities in addressing scientific questions that could not be answered otherwise. Thus, we are confident in a prompt integration of these methods into astronomy, with the hope that contemporary statistical techniques may become common practice in the 21st century astrophysical research.

Acknowledgements

MK acknowledges support by the DFG project DO 1310/4-1. UM would like to thank funding from a Marie Curie Fellowship of the European Union Seventh Framework Project (FP7/2007-2013), grant agreement n. 267251. Work on this paper has substantially benefited from using the collaborative website AWOB (http://awob.mpg.de) developed and maintained by the Max-Planck Institute for Astrophysics and the Max-Planck Digital Library. The bibliographic research was possible thanks to the tools offered by the NASA Astrophysical Data Systems and the JSTOR archive.

Appendix A. Flowchart for GLM regression

This section illustrates a brief summary of GLM analysis and model diagnostics. It comprises:

- Acquire the dataset.
- Choose the response variable to be modelled.
- Choose predictor variables.
- Choose GLM family, e.g. Gaussian, Poisson, binomial.
- Choose either a maximum-likelihood or a Bayesian approach.
- Choose link function.
- Estimating coefficients by means of a GLM or Bayesian GLM analysis, i.e., estimate $\eta$ and predicted probabilities $\pi$.
- Classification and diagnostic tests:
  - ROC curve-probability threshold.
  - Confusion Matrix for a given $\pi_{\text{ish}}$ and assigned class memberships.

References

Akaike, H., December 1974. A new look at the statistical model identification. IEEE Transactions on Automatic Control 19 (6), 716–723.
Albert, J., 2007. Bayesian computation with R. Vol. 747389981. Springer.
Albert, J. H., Chib, S., 1993. Bayesian analysis of binary and polychotomous response data. Journal of the American Statistical Association 88 (422), 669–679.
Figure A.8: Tabular data is represented by blue rectangles, calculations by red diamonds, choices by green parallelograms, and diagnostic outcomes by orange heptagons.
Andreon, S., Hurn, M. A., Jun. 2010. The scaling relation between richness and mass of galaxy clusters: a Bayesian approach. MNRAS 404, 1922–1937.

Ata, M., Kitaura, F.-S., Müller, V., Aug. 2014. Bayesian inference of cosmic density fields from non-linear, scale-dependent, and stochastic biased tracers. astro/1408.2566.

Bate, M. R., Burkert, A., Jul. 1997. Resolution requirements for smoothed particle hydrodynamics calculations with self-gravity. MNRAS 288, 1060–1072.

Biffi, V., Maio, U., Dec. 2013. Statistical properties of mass, star formation, chemical content and rotational patterns in early z ≥ 9 structures. MNRAS 436, 1621–1638.

Bishop, C. M., 1996. Neural Networks for Pattern Recognition. Oxford University Press, 1st Edition.

Bornkamp, B., 2012. Functional uniform priors for nonlinear modeling. Biometrics 68 (3), 893–901.

Breslow, N. E., Clayton, D. G., 1993. Approximate Inference in Generalized Linear Mixed Models. Journal of the American Statistical Association 88 (421), 9–25.

Bromm, V., Nov. 2013. Formation of the first stars. Reports on Progress in Physics 76 (11), 112901.

Bromm, V., Ferrara, A., Coppi, P. S., Larson, R. B., Dec. 2001. The fragmentation of pre-enriched primordial objects. MNRAS 328, 969–976.

Cameron, E., 2011. On the estimation of confidence intervals for binomial population proportions in astronomy: The simplicity and superiority of the bayesian approach. Publications of the Astronomical Society of Australia 28, 128–139.

de Jong, P., Heller, G. Z., 2008. Generalized Linear Models for Insurance Data. Cambridge University Press, cambridge Books Online. URL http://dx.doi.org/10.1017/CBO9780511755408

de Souza, R. S., Ciardi, B., Maio, U., Ferrara, A., Jan. 2013a. Dark matter halo environment for primordial star formation. MNRAS 428, 2109–2117.

de Souza, R. S., Ishida, E. E. O., Johnson, J. L., Whalen, D. J., Mesinger, A., Dec. 2013b. Detectability of the first cosmic explosions. MNRAS 436, 1555–1563.

de Souza, R. S., Krone-Martins, A., Ishida, E. E. O., Ciardi, B., Sep. 2012. Searching for the first stars with the Gaia mission. A&A545, A102.

de Souza, R. S., Maio, U., Biffi, V., Ciardi, B., May 2014. Robust PCA and MIC statistics of baryons in early minihaloes. MNRAS 440, 240–248.

de Souza, R. S., Yoshida, N., Ioka, K., Sep. 2011. Populations III.1 and III.2 gamma-ray bursts: constraints on the event rate for future radio and X-ray surveys. A&A533, A32.

Diggle, P., Moyeed, R., Rowlingson, B., Thomson, M., 2002. Childhood malaria in the gambia: a case-study in model-based geostatistics. Journal of the Royal Statistical Society: Series C (Applied Statistics) 51 (4), 493–506.

Dolag, K., Borgani, S., Murante, G., Springel, V., Oct. 2009. Substructures in hydrodynamical cluster simulations. MNRAS 399, 497–514.

Doss, H., Narasimhan, B., 1994. Bayesian poisson regression using the gibbs sampler: Sensitivity analysis through dynamic graphics. Tech. rep., Technical report, Penn State Erie.

Duda, R., Hart, P. E., Stork, D. G., 2000. Pattern Classification. Wiley-Interscience, 2nd. Edition.

Gelman, A., Jakulin, A., Pittau, M. G., Su, Y.-S., 2008. A weakly informative default prior distribution for logistic and other regression models. The Annals of Applied Statistics, 1360–1383.

Gelman, A., Su, Y.-S., 2014. arm: Data Analysis Using Regression and Multilevel/Hierarchical Models. R package version 1.7-05. URL http://CRAN.R-project.org/package=arm

Gini, C., 1921. Variabilità e mutabilità. Memorie di metodologica statistica (Ed. Pizetti E, Salvemini, T).

Gini, C., 1921. Measurement of inequality of incomes. The Economic Journal 31 (121), 124–126.

Grisswold, M., Parmigiani, G., Potosky, A., Lipscomb, J., 2004. Analyzing health care costs: a comparison of statistical methods motivated by medicare colorectal cancer charges. Biostatistics 1 (1), 1–23.

Hardin, J. W., Hilbe, J. M., 2012. Generalized Linear Models and Extensions, 3rd Edition, 3rd Edition. StataCorp LP.

Hastie, T., Tibshirani, R., 1986. Generalized additive models. Statistical science, 297–310.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, 2nd. Edition.

Heger, A., Woosley, S. E., Mar. 2002. The Nucleosynthetic Signature of Population III. ApJ567, 532–543.

Hilbe, J., 2009. Logistic Regression Models. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis. URL http://books.google.es/books?id=eJcHIAAAACAAJ

Hilbe, J., 2014. Modeling Count Data. Cambridge University Press.

Hubble, E., 1929. A relation between distance and radial velocity among extra-galactic nebulae. Proceedings of the National Academy of Sciences 15 (3), 168–173.

Ibrahim, J. G., Laud, P. W., 1991. On bayesian analysis of generalized linear models using Jeffreys’s prior. Journal of the American Statistical Association 86 (416), pp. 981–986. URL http://www.jstor.org/stable/2290514

Isobe, T., Feigelson, E. D., Akritas, M. G., Babu, G. J., 1990. Linear regression in astronomy. The Astrophysical Journal 364, 104–113.

Kohavi, R., Provost, F., 1998. Glossary of terms. Machine Learning 30 (2-3), 271–274.

Lansbury, G. B., Lucey, J. R., Smith, R. J., 2014. Barred s0 galaxies in the coma cluster. Monthly Notices of the Royal Astronomical Society 439 (2), 1749–1764.

Liang, K.-Y., Zeger, S. L., 1986. Longitudinal data analysis using generalized linear models. Biometrika 73 (1), 13–22.

Lindsey, J. K., 1999. A review of some extensions to generalized linear models. Statistics in medicine 18 (17-18), 2223–2236.
