Computational Statistics and Data Analysis 121 (2018) 71–89

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Efficient estimation of COM–Poisson regression and a generalized additive model

Suneel Babu Chatla *, Galit Shmueli

Institute of Service Science, National Tsing Hua University, Hsinchu 30013, Taiwan

ARTICLE INFO

Article history:
Received 17 April 2017
Received in revised form 1 November 2017
Accepted 16 November 2017
Available online 20 December 2017

Keywords:
IRLS
Penalized splines
P-IRLS
Over and under dispersion
Time series

ABSTRACT

The Conway–Maxwell–Poisson (CMP) or COM–Poisson regression is a popular model for count data due to its ability to capture both under dispersion and over dispersion. However, CMP regression is limited when dealing with complex nonlinear relationships. With today's wide availability of count data, especially due to the growing collection of data on human and social behavior, there is need for count data models that can capture complex nonlinear relationships. One useful approach is additive models; but, there has been no additive model implementation for the CMP distribution. To fill this void, we first propose a flexible estimation framework for CMP regression based on iterative reweighted least squares (IRLS) and then extend this model to allow for additive components using a penalized splines approach. Because the CMP distribution belongs to the exponential family, convergence of IRLS is guaranteed under some regularity conditions. Further, it is also known that IRLS provides smaller standard errors compared to gradient-based methods. We illustrate the usefulness of this approach through extensive simulation studies and using real data from a bike sharing system in Washington, DC.

1. Introduction

Count data have become popular dependent variables in studies in various areas, especially due to the growing availability of data on human and social behavior. Examples include the number of crimes in each neighborhood, number of accidents at an intersection, number of Facebook comments, ridership in bike sharing programs, etc. The wide availability of count data and the need for modeling such data as a function of other factors to establish causal relationships or to quantify correlated relationships has led to the widespread use of count data models.

The most commonly used regression models for cross-sectional count data are Poisson regression and Negative-Binomial regression. In addition, the Conway–Maxwell–Poisson (CMP) distribution (also known as the COM–Poisson distribution) has gained increasing popularity for its flexibility and ability to handle both over and under dispersed data. Revived by Shmueli et al. (2005), the CMP distribution is a two-parameter generalization of the Poisson, Bernoulli, and Geometric distributions. Suppose Y is a random variable that follows a CMP distribution, then the probability mass function (p.m.f.) for $Y \in \{0, 1, 2, \ldots\}$ is defined as

$$P(Y = y) = \frac{\lambda^y}{(y!)^\nu \zeta(\lambda, \nu)}, \quad \text{where} \quad \zeta(\lambda, \nu) = \sum_{s=0}^{\infty} \frac{\lambda^s}{(s!)^\nu}$$

for the parameters $\lambda, \nu > 0$ and $0 < \lambda < 1, \nu = 0$.

* Corresponding author.
E-mail addresses: suneel.chatla@iss.nthu.edu.tw (S.B. Chatla), galit.shmueli@iss.nthu.edu.tw (G. Shmueli).

https://doi.org/10.1016/j.csda.2017.11.011
0167-9473/© 2017 Elsevier B.V. All rights reserved.

The CMP distribution includes three well-known distributions as special cases: Poisson ($\nu = 1$), Geometric ($\nu = 0, \lambda < 1$), and Bernoulli ($\nu \to \infty$ with probability $\frac{\lambda}{1+\lambda}$). Due to the additional parameter ν, the CMP distribution is flexible enough to handle both over dispersion ($\nu < 1$) and under dispersion ($\nu > 1$) which are common in count data (Sellers and Shmueli, 2010). For more details on the distributional properties please refer to Daly and Gaunt (2016), Sellers et al. (2012), Minka et al. (2003).

One of the major limitations of the CMP distribution is that the normalizing constant $\zeta(\lambda, \nu)$, which is an infinite series, does not have a closed form representation, and therefore there is no closed form representation available for the mean. This makes it difficult to model the mean directly as a function of covariates, as in standard models such as Poisson and Logistic regression. However, the CMP distribution belongs to the exponential family and thus has the properties and advantages of that family. Defining $\boldsymbol{\theta} = (\lambda, \nu)$, the CMP likelihood has the following form of an exponential family (Lehmann and Casella, 2006):

$$P_Y(y|\boldsymbol{\theta}) = h(y)\exp\left(\sum_{i=1}^{2} \eta_i(\boldsymbol{\theta})T_i(y) - A(\boldsymbol{\theta})\right),$$

where the natural parameters are $\eta_1(\boldsymbol{\theta}) = \ln\lambda$ and $\eta_2(\boldsymbol{\theta}) = -\nu$ with corresponding sufficient statistics $T_1(y) = y$, $T_2(y) = \ln(y!)$ and $A(\boldsymbol{\theta}) = \ln\zeta(\lambda, \nu)$, $h(y) = 1$, as mentioned in Shmueli et al. (2005).

Although CMP regression is flexible in terms of handling both over and under dispersion, it is sometimes too restrictive for modeling nonlinear relationships or time series data. At the same time, additive models are widely used for modeling nonlinear relationships such as time series (Dominici et al., 2002; Stieb et al., 2003). Additive models have the advantage of being parsimonious while at the same time providing more flexibility to capture complicated relationships. Currently, there exists no additive model implementation for the CMP regression. Motivated by the need for flexible count data regression models for applications such as bike sharing, which can assist service providers in better management of their resources, we develop an additive model for CMP regression. Existing additive model implementations are heavily dependent upon the iterative reweighted least squares (IRLS) estimation framework, which currently does not exist for CMP regression. In this study, we propose and implement an IRLS estimation framework for CMP regression and then extend that to additive models.

The outline of this paper is as follows: In Section 2, we describe the CMP regression and the problems associated with IRLS implementation. In Section 3, we develop an IRLS framework for estimating a CMP regression by providing theory and the pseudo algorithm. We compare our proposed IRLS methodology with existing methods using an extensive simulation study in Section 4. In Section 5, we use the IRLS framework to develop an additive model for the CMP distribution, and again evaluate its performance using a simulation in Section 6. In Section 7, we use our proposed additive model to draw valuable insights from a bike sharing application. Section 8 presents conclusions and future directions.

2. CMP regression

Assume that we have a random sample of n observations $\{y_i, \boldsymbol{x}_i^T, \boldsymbol{z}_i^T\}_{i=1}^n$, where $\boldsymbol{x}_i^T = [1, x_{i1}, \ldots, x_{ip}]$ and $\boldsymbol{z}_i^T = [1, z_{i1}, \ldots, z_{iq}]$. In matrix notation, let $Y = [y_1, \ldots, y_n]^T$, $X = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n]^T$ and $Z = [\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n]^T$ with the parameter vectors $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_n)^T$, $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_n)^T$ and $\boldsymbol{\zeta} = (\zeta_1, \ldots, \zeta_n)^T$. We also denote mean and variance functions as $E[\cdot]$, $V[\cdot]$ respectively.

When needed, we use the vector notation. With a slight abuse of notation, we extend the operations on scalars to operations on vectors. For example, we write $\ln(Y!) = (\ln(y_1!), \ldots, \ln(y_n!))^T$, $\ln(\boldsymbol{\lambda}) = (\ln(\lambda_1), \ldots, \ln(\lambda_n))^T$ and $\frac{\partial \ln \boldsymbol{\zeta}}{\partial \ln \boldsymbol{\lambda}} = (\frac{\partial \ln \zeta_1}{\partial \ln \lambda_1}, \ldots, \frac{\partial \ln \zeta_n}{\partial \ln \lambda_n})^T$. Unless otherwise stated, any operation on a vector simply denotes an extension of that operation to each component of that vector.

The CMP regression can be formulated as

$$\ln(\boldsymbol{\lambda}) = X\boldsymbol{\beta} \tag{1}$$

$$\ln(\boldsymbol{\nu}) = Z\boldsymbol{\gamma} \tag{2}$$

where $\boldsymbol{\beta} \in \mathbb{R}^{p+1}$, $\boldsymbol{\gamma} \in \mathbb{R}^{q+1}$.

The log link is used for the $\boldsymbol{\lambda}$ model. As mentioned in Sellers and Shmueli (2010), this choice of log link is useful for two reasons. First, it coincides with the link function in two well-known cases: in Poisson regression, it reduces to $E[y_i] = \lambda_i$; in logistic regression, where $p_i = \frac{\lambda_i}{1+\lambda_i}$, it reduces to $logit(p_i) = \ln\lambda_i$. The second advantage of using a log link function is that it leads to elegant estimation, inference, and diagnostics. At the same time, we deliberately consider a log link for the $\boldsymbol{\nu}$ model, although the canonical link is identity, to restrict model predictions to the range $(0, \infty)$. This is important because while $\boldsymbol{\gamma}$ is unconstrained, ν can only take non-negative values and we cannot use the identity link between $\boldsymbol{\nu}$ and $\boldsymbol{\gamma}$.

In applications, it is common to treat $\boldsymbol{\nu}$ as a nuisance parameter. For this reason, usually the data matrix Z contains only the intercept. Yet, since the $\boldsymbol{\nu}$ parameter models the dispersion, it is always better to include covariates that can potentially control for it (Sellers and Shmueli, 2013). In theory, one could use the same predictors for modeling both parameters. However, in practice, to avoid collinearity issues, it is better to have at least one different covariate in either the $\ln(\boldsymbol{\lambda})$ or the $\ln(\boldsymbol{\nu})$ model.