Applied Mathematical Modelling 56 (2018) 172–185

Contents lists available at ScienceDirect

Applied Mathematical Modelling

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

On the Bell distribution and its associated regression model for count data

Fredy Castellares[a], Silvia L.P. Ferrari[b,*], Artur J. Lemonte[c]

[a] Departamento de Estatística, Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brazil
[b] Departamento de Estatística, Universidade de São Paulo, São Paulo, SP, Brazil
[c] Departamento de Estatística, Universidade Federal do Rio Grande no Norte, Natal, RN, Brazil

A R T I C L E   I N F O

Article history:
Received 13 July 2017
Revised 26 October 2017
Accepted 4 December 2017
Available online 8 December 2017

Keywords:
Count data
Multiple Poisson process
Overdispersion
Poisson distribution

A B S T R A C T

In this paper we define and study the one-parameter discrete Bell distribution, which has a very simple form for its probability mass function. In particular, we show that this discrete distribution is a particular solution of a multiple Poisson process. By considering the distribution studied in this article, we introduce a new regression model where the response variable is a count. Further, residuals are also proposed for the new count regression model. An empirical application is considered to show the usefulness of the Bell regression model in practice.

1. Introduction

Count data occur in many practical problems as, for example, the number of occurrences of thunderstorms in a calendar year, the number of accidents, the number of absences at work, the number of days lost due to work accidents, the number of insurance claims, the number of species in a habitat, and so on. Undoubtedly, the one-parameter Poisson distribution is the most popular model for count data used in practice, mainly because of its simplicity. A major drawback of this distribution is that the variance is restricted to be equal to the mean. So, alternative discrete probability distributions, which describe count phenomena, have been proposed in the statistical literature, due perhaps to advances in computational methods which enable us to compute, straightforwardly, the numerical value of special functions such as hypergeometric series. It is worth emphasizing that most of the new discrete distributions are obtained by discretization of a known continuous distribution, and have more than one parameter.

In this paper, instead of discretizing a known continuous distribution, we shall introduce a new one-parameter discrete family of distributions on the basis of a series expansion due to Bell [1,2] the so-called *Bell distribution*, which is very simple to deal with, since its probability mass function does not contain any complicated function. The Bell distribution has many interesting properties such as (among many others): (i) it is an one-parameter distribution; (ii) it belongs to the one-parameter exponential family of distributions; (iii) the Poisson distribution is not nested in the Bell family, but for small values of the parameter the Bell distribution approaches the Poisson distribution; and (iv) it is infinitely divisible. We also derive some mathematical properties of the Bell family of distributions. In particular, we show that the Bell distribution is a special case of a multiple Poisson process.

* Corresponding author.
E-mail address: silviaferrari@usp.br (S.L.P. Ferrari).

https://doi.org/10.1016/j.apm.2017.12.014

On the basis of the Bell distribution, we also propose a new regression model where the response variable is a count. Similarly to the generalized linear model setup, the mean response of the new regression model is related to a linear predictor through a link function, which allows for parameter interpretation in terms of the response variable in the original scale. Furthermore, the new regression model does not depend on complicated functions, and some quantities (e.g., score function, Fisher information matrix, etc.) related to the Bell regression model are simple and compact. We also consider two residuals for this class of regression models, which are simple enough for practical purposes and follow the lines of the generalized linear models framework. We illustrate the Bell regression model in a real data application and verify that this regression model may be a useful alternative to the usual Poisson and negative binomial (NB) regression models.

The rest of this article is organized as follows. In Section 2 we introduce the Bell distribution, and derive several structural properties related to this distribution. Section 3 deals with the Bell distribution parameter estimation. The Bell regression model is introduced in Section 4. Residuals are considered in Section 5. Section 6 contains an application to real data of the Bell regression model for illustrative purposes. Section 7 closes up the paper with some concluding remarks.

2. The Bell distribution and some of its properties

In order to introduce the one-parameter Bell distribution, we consider the following expansion provided by Bell [1,2]:

$$\exp(e^x - 1) = \sum_{n=0}^{\infty} \frac{B_n}{n!} x^n, \quad x \in \mathbb{R}, \tag{1}$$

where the coefficients B_n are the Bell numbers defined by

$$B_n = \frac{1}{e} \sum_{k=0}^{\infty} \frac{k^n}{k!}. \tag{2}$$

Starting with $B_0 = B_1 = 1$, the first few Bell numbers are $B_2 = 2$, $B_3 = 5$, $B_4 = 15$, $B_5 = 52$, $B_6 = 203$, $B_7 = 877$, $B_8 = 4140$, $B_9 = 21147$, $B_{10} = 115975$, $B_{11} = 678570$, $B_{12} = 4213597$ and $B_{13} = 27644437$.

Remark 1. The Bell number B_n in Eq. (2) is the nth moment of the Poisson distribution with parameter equal to 1.

By considering (1) and (2), we define the Bell distribution.

Definition 1. A discrete random variable Y has a Bell distribution with parameter θ if its probability mass function is given by

$$\Pr(Y = y) = \frac{\theta^y \, e^{-e^\theta + 1} B_y}{y!}, \quad y = 0, 1, 2, \ldots, \tag{3}$$

where $\theta > 0$, and B_y are the Bell numbers in (2).

If Y follows a Bell distribution with parameter $\theta > 0$, then the notation used is $Y \sim \mathrm{Bell}(\theta)$. The Bell probability mass function in (3) is very simple to deal with, and it does not involve any complicated function. For example, we have:

$$\Pr(Y = 0) = e^{-e^\theta + 1}, \quad \Pr(Y = 1) = \theta \, e^{-e^\theta + 1},$$

$$\Pr(Y = 2) = \theta^2 e^{-e^\theta + 1}, \quad \Pr(Y = 3) = \frac{5\,\theta^3}{3!} e^{-e^\theta + 1},$$

$$\Pr(Y = 4) = \frac{15\,\theta^4}{4!} e^{-e^\theta + 1}, \quad \Pr(Y = 5) = \frac{52\,\theta^4}{5!} e^{-e^\theta + 1},$$

and the other probabilities can also be easily obtained. Fig. 1 displays graphics of the Bell probability mass function for different values of θ.

Theorem 1 below, due to Jánossy et al. [3], will be used to show that the Bell distribution is a special case of a multiple Poisson process.

Theorem 1. *Let us consider events occurring in time and let us impose the following conditions:*

(a) *The process is homogeneous in time, i.e., we assume that the probability of exactly k events occurring in the interval (t_1, t_2) depends only on the length $t = t_2 - t_1$ of this interval; let this probability be denoted by $W_k(t)$, for $k = 0, 1, 2, \ldots$. Evidently we have*

$$W_k(t) \geq 0 \quad \text{and} \quad \sum_{k=0}^{\infty} W_k(t) = 1 \quad \text{for any} \quad t \geq 0.$$

Further,

$$W_0(0) = 1 \quad \text{and thus} \quad W_k(0) = 0 \quad \text{for} \quad k = 1, 2, \ldots.$$

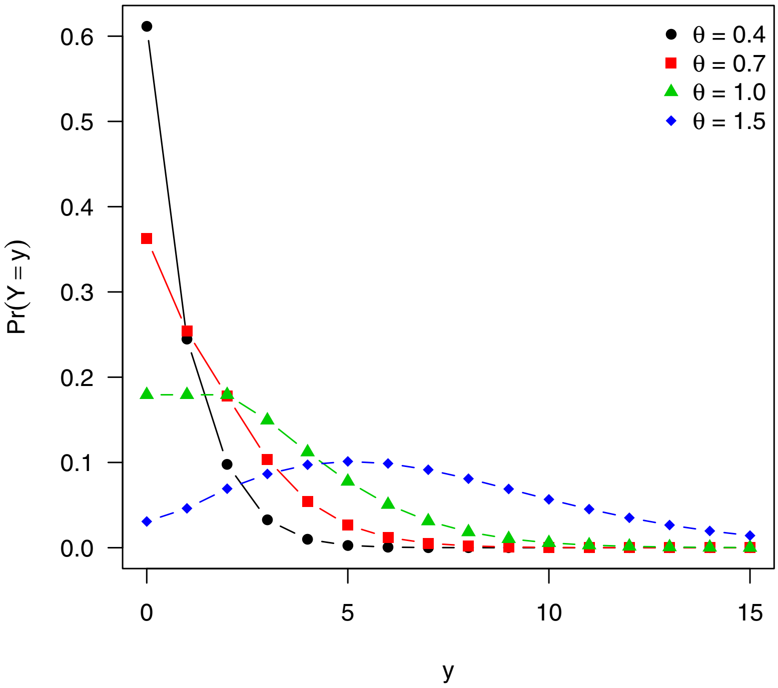

Fig. 1. The Bell probability mass function for different values of θ.

(b) *The process is Markov's type, i.e., the number of events occurring during the time interval* (t_1, t_2) *is independent of the numbers of events occurring during the time interval* (t_3, t_4) *provided that* $t_1 < t_2 \le t_3 < t_4$.

Then,

$$W_k(t) = \exp\left(-t\sum_{n=1}^{\infty} c_n\right) \sum_{r_1+2r_2+\cdots+kr_k=k} \frac{(c_1 t)^{r_1}(c_2 t)^{r_2}\cdots(c_k t)^{r_k}}{r_1!\,r_2!\cdots r_k!}, \tag{4}$$

for a sequence of constants $\{c_n,\ n \ge 1\}$ *such that* $c_n \ge 0$ *and* $\sum_{n=1}^{\infty} c_n$ *is convergent.*

Proof. The reader is referred to Jánossy et al. [3]. □

Remark 2. A multiple Poisson process is a Markovian process with independent increments where the probability of occurrence of more than one event in a short period $(0, t)$ is non negligible relatively to the probability of the occurrence of a single event. Unlike the Poisson processes, it allows the occurrence of multiple events. The difference between this process and the Poisson process can be found in Jánossy et al. [3]. Under conditions (a) and (b) in Theorem 1 and also the "rare event" condition:

$$\lim_{t\to 0} \frac{W_1(t)}{1 - W_0(t)} = 1,$$

the resulting process is a Poisson process. The rare event condition means that in a short interval the probability of the occurrence of two or more events becomes arbitrarily small when compared with the probability of the occurrence of exactly one event in the same time interval.

We have the following corollary.

Corollary 1. *Let* $\{c_n = \theta^n/n!,\ n \ge 1\}$ *be a sequence of constants for each* $\theta > 0$. *For the time interval length* $t = 1$, *we have*

$$W_y(1) := W_y = \frac{\theta^y\, e^{-e^{\theta}+1} B_y}{y!}, \quad y = 0, 1, 2, \ldots,$$

where B_y *are the Bell numbers in (2), i.e.* W_y *is the probability mass function of the Bell distribution with parameter* θ.

Proof. Let $\theta > 0$, $c_n = \theta^n/n!$, for $n = 1, 2, \ldots,$ and $t = 1$. From (4), it follows that

$$W_k(1) = \exp\left(-\sum_{n=1}^{\infty} \frac{\theta^n}{n!}\right) \sum_{r_1+2r_2+\cdots+kr_k=k} \frac{\theta^{r_1}\theta^{2r_2}\cdots\theta^{kr_k}}{(1!)^{r_1}(2!)^{r_2}\cdots(k!)^{r_k}\, r_1!\, r_2!\cdots r_k!}.$$

We have that

$$W_k(1) = \exp(1 - e^{\theta}) \sum_{r_1+2r_2+\cdots+kr_k=k} \frac{\theta^{r_1+2r_2+\cdots+kr_k}}{(1!)^{r_1}(2!)^{r_2}\cdots(k!)^{r_k} r_1!\, r_2! \cdots r_k!}.$$

By using $r_1 + 2r_2 + \cdots + kr_k = k$ in the above expression, it then follows that

$$W_k(1) = \theta^k \exp(1 - e^{\theta}) \sum_{r_1+2r_2+\cdots+kr_k=k} \frac{1}{(1!)^{r_1}(2!)^{r_2}\cdots(k!)^{r_k} r_1!\, r_2! \cdots r_k!}.$$

By using Eq. (4.5) of Bell [1, p. 264], the above sum reduces to

$$\sum_{r_1+2r_2+\cdots+kr_k=k} \frac{1}{(1!)^{r_1}(2!)^{r_2}\cdots(k!)^{r_k} r_1!\, r_2! \cdots r_k!} = \frac{B_k}{k!},$$

where B_k are the Bell numbers in (2). This completes the proof. □

Remark 3. Corollary 1 allows us to associate a count distribution given by the Bell distribution with parameter $\theta > 0$ to the distribution of a multiple Poisson process (with rates $c_n = \theta^n/n!$).

In what follows, we present some results related to the Bell distribution. (All proofs are omitted to save space, but may be requested from the authors.)

Proposition 1. *Let $Y \sim$ Bell(θ), where $\theta > 0$. Then, the probability generating function is*

$$G_Y(s) = \mathbb{E}(s^Y) = \exp(e^{s\theta} - e^{\theta}), \quad |s| < 1.$$

Remark 4. The mean and variance of $Y \sim \text{Bell}(\theta)$ are

$$\mathbb{E}(Y) = \theta\, e^{\theta}, \quad \mathbb{VAR}(Y) = \theta\,(1+\theta)\, e^{\theta}.$$

Remark 5. The index of dispersion (a normalized measure of the dispersion of a probability distribution) of the Bell distribution, defined as $I_d = \mathbb{VAR}(Y)/\mathbb{E}(Y)$, takes the form $I_d = 1 + \theta$. It follows that $I_d > 1$ for all $\theta > 0$ and hence the Bell distribution may by suitable for modeling count data with overdispersion. Clearly, the overdispersion is of a specific form and is constrained by the mean. Therefore, the Bell distribution may not accommodate all possible forms of overdispersion.

Proposition 2. *Let $Y \sim$ Bell(θ), where $\theta > 0$. Then, we have that:*

(a) *The Kullback–Leibler divergence between the Bell(θ) and Poisson(λ) distributions takes the form*

$$D_{\text{KL}}(\lambda, \theta) = e^{\theta} - 1 - \lambda + \lambda \log(\lambda/\theta) - e^{-\lambda} \sum_{n=0}^{\infty} \frac{\log(B_n)}{n!} \lambda^n,$$

where B_n are the Bell numbers in (2).

(b) *The Kullback–Leibler entropy between the Bell(θ) and Bell(φ) distributions becomes*

$$I(\theta, \varphi) = \theta\, e^{\theta} \log(\theta/\varphi) + e^{\varphi} - e^{\theta}.$$

Proposition 3. *Let $Y \sim$ Bell(θ), where $\theta > 0$. Then, the random variable Y has the same distribution as the sum of N independent and identically zero-truncated Poisson distributed random variables with parameters $\theta > 0$, with $N \sim$ Poisson($e^{\theta} - 1$).*

Remark 6. Proposition 3 allows us to provide the following characterization. Let $X_1, X_2, X_3, \ldots$ be a sequence of independent and identically distributed random variables such that X_n has zero-truncated Poisson distribution with parameter $\theta > 0$, and let N be Poisson distributed with parameter $e^{\theta} - 1$ and independent of the sequence $\{X_n,\ n \geq 1\}$. Then, the random variable Y given by

$$Y = X_1 + X_2 + \cdots + X_N$$

has a Bell(θ) distribution, where $\theta > 0$.

Proposition 4. *The Bell distribution is identifiable for all $\theta > 0$.*

Proposition 5. *The Bell distribution is strongly unimodal.*

Proposition 6. *The Bell distribution is infinitely divisible.*

Remark 7. Infinitely divisible distributions play an important role in many areas of statistics, for example, in stochastic processes and in actuarial statistics. When a distribution G is infinitely divisible, then for any integer $j \geq 2$, there exists a distribution G_j such that G is the j-fold convolution of G_j, namely, $G = G_j^{*j}$.

Proposition 7. *Let $Y_1, Y_2, \ldots, Y_n$ be n independent and identically distributed random variables such that $Y_i \sim$ Bell(θ) for $i = 1, 2, \ldots, n$. Then, the probability mass function of $S_n = Y_1 + Y_2 + \cdots + Y_n$ is given by*

$$\Pr(S_n = x) = \frac{e^{n(-e^{\theta}+1)}\, \theta^x\, T_x(n)}{x!}, \qquad x = 0, 1, 2, \ldots,$$

where $\theta > 0$, and $T_x(n)$are Touchard polynomials [4]. The first few Touchard polynomials are $T_0(n) = 1,\ T_1(n) = n,\ T_2(n) = n^2 + n,\ T_3(n) = n^3 + 3n^2 + n,\ T_4(n) = n^4 + 6n^3 + 7n^2 + n,\ T_5(n) = n^5 + 10n^4 + 25n^3 + 15n^2 + n,\ T_6(n) = n^6 + 15n^5 + 65n^4 + 90n^3 + 31n^2 + n$, etc.

Remark 8. Proposition 7 shows that the distribution of a sum of n independent random variables having a Bell(θ) distribution is the Neyman type A distribution with parameters θ and ne^{θ}. This result may be used to evaluate probabilities of events involving the sample mean of Bell *i.i.d.* random variables.

3. Parameter estimation

If $Y \sim$ Bell(θ), then its probability mass function can be expressed as

$$\Pr(Y = y) = \exp\left(y\log(\theta) + \log(B_y/y!) - e^{\theta} + 1\right) = \exp\left(\xi\, T(y) - A(\theta) + C(y)\right),$$

where $y = 0, 1, 2, \ldots,$ and hence the Bell distribution belongs to the one-parameter exponential family of distributions with natural parameter $\xi = \log(\theta),\ T(y) = y,\ A(\theta) = e^{\theta}$ and $C(y) = \log(B_y/y!) + 1$. Additionally, if $Y_1, Y_2, \ldots, Y_n$ is a random sample of size n from $Y \sim$ Bell(θ), then $T(Y_1, Y_2, \ldots, Y_n) = \sum_{i=1}^n Y_i$ is a complete sufficient statistic for θ. The log-likelihood function based on the observed sample $y_1, y_2, \ldots, y_n$, apart from constants, is

$$\ell(\theta) = -n\, e^{\theta} + \log(\theta)\sum_{i=1}^{n} y_i.$$

The maximum likelihood (ML) estimator $\widehat{\theta}$ of θ satisfies the equation

$$-e^{\widehat{\theta}} + \frac{\overline{Y}}{\widehat{\theta}} = 0;$$

or, equivalently, $\overline{Y} = \widehat{\theta}\exp(\widehat{\theta})$, whose solution is $\widehat{\theta} = W_0(\overline{Y})$, where $W_0(\cdot)$ is the Lambert function (Corless et al. [5]), and $\overline{Y} = n^{-1}\sum_{i=1}^n Y_i$. Also, the second derivative of $\ell(\theta)$ in relation to θ evaluated at the ML estimate $\widehat{\theta} = W_0(\overline{y})$ with $\overline{y} = n^{-1}\sum_{i=1}^n y_i$ is negative; that is,

$$\left.\frac{d^2\ell(\theta)}{d\theta^2}\right|_{\theta=\widehat{\theta}} = -\left[n\, e^{W_0(\overline{y})} + \frac{n\,\overline{y}}{W_0(\overline{y})^2}\right] < 0.$$

It is easy to verify that the Fisher information for θ is simply given by $n\theta^{-1}(1+\theta)e^{\theta}$. Finally, when n is large and under some mild regularity conditions, we have that $\sqrt{n}(\widehat{\theta} - \theta) \overset{a}{\sim} \mathcal{N}(0, \theta(1+\theta)^{-1}e^{-\theta})$, where "$\overset{a}{\sim}$" means approximately distributed.

4. The regression model

The Bell probability mass function is given in (3), where it is indexed by the parameter $\theta > 0$. Its mean and variance are, respectively,

$$\mathbb{E}(Y) = \theta\, e^{\theta}, \tag{5}$$

$$\mathbb{VAR}(Y) = \theta\,(1+\theta)\, e^{\theta}. \tag{6}$$

In a regression model framework, it is typically more useful to model the mean of the response variable. So, to obtain a regression structure for the mean of the Bell distribution, we shall work with a different parameterization of the Bell mass probability function. Let $\mu = \theta\, e^{\theta}$ and hence $\theta = W_0(\mu)$, where $W_0(\cdot)$ is the Lambert function. It then follows from (5) and (6) that

$$\mathbb{E}(Y) = \mu, \quad \mathbb{VAR}(Y) = \mu[1 + W_0(\mu)],$$

so that $\mu > 0$ is the mean of the response variable Y. The Bell probability mass function can be written, in the new parameterization, as

$$\Pr(Y = y) = \exp\left(1 - e^{W_0(\mu)}\right)\frac{W_0(\mu)^y B_y}{y!}, \quad y = 0, 1, 2, \ldots, \tag{7}$$

where $\mu > 0$, and B_y are the Bell numbers in (2). As early noted, the probability mass function (7) belongs to the one-parameter exponential family. Then, the variance function can be easily obtained, which is given by $V(\mu) = \mu[1 + W_0(\mu)]$.

We have that $W_0(\mu) > 0$ for $\mu > 0$ and, therefore, $\mathbb{VAR}(Y) > \mathbb{E}(Y)$. It implies that the Bell distribution may be suitable for modeling count data with overdispersion, like the two-parameter NB distribution. An advantage of the Bell distribution in relation to the NB distribution is that no additional (dispersion) parameter is necessary to accommodate overdispersion.

Let $Y_1, Y_2, \ldots, Y_n$ be n independent random variables, where each Y_i, for $i = 1, 2, \ldots, n$, follows the probability mass function (7) with mean μ_i; that is, $Y_i \sim \mathrm{Bell}(W_0(\mu_i))$, for $i = 1, 2, \ldots, n$. Suppose the mean of Y_i satisfies the following functional relation:

$$g(\mu_i) = \eta_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}, \quad i = 1, 2, \ldots, n, \tag{8}$$

where $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^\top \in \mathbb{R}^p$ is a p-dimensional vector of regression coefficients ($p < n$), η_i is the linear predictor, and $\boldsymbol{x}_i^\top = (x_{i1}, x_{i2}, \ldots, x_{ip})$ denotes the observations on p known covariates. Note that the variance of Y_i is a function of μ_i and, as a consequence, of the covariate values. Hence, non-constant response variances are naturally accommodated into the model. We assume that the mean link function $g: (0, \infty) \to \mathbb{R}$ is strictly monotonic and twice differentiable. There are several possible choices for the mean link function. For instance, some useful link functions are: logarithmic $g(\mu) = \log(\mu)$, square root $g(\mu) = \sqrt{\mu}$ and identity $g(\mu) = \mu$ (with special attention on the positivity of the estimates), among others; see also McCullagh and Nelder [6].

The ML method is considered to estimate the parameter vector $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^\top$. The log-likelihood function, except for constant terms, is given by

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left[y_i \log(W_0(\mu_i)) - \mathrm{e}^{W_0(\mu_i)} \right],$$

where $\mu_i = g^{-1}(\eta_i)$ is a function of $\boldsymbol{\beta}$, and $g^{-1}(\cdot)$ is the inverse of $g(\cdot)$. The score function is given by the p-vector

$$\boldsymbol{U}(\boldsymbol{\beta}) = \boldsymbol{X}^\top \boldsymbol{W}^{1/2} \boldsymbol{V}^{-1/2} (\boldsymbol{y} - \boldsymbol{\mu}),$$

where the model matrix $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)^\top$ has full column rank, $\boldsymbol{W} = \mathrm{diag}\{w_1, w_2, \ldots, w_n\}$, $\boldsymbol{V} = \mathrm{diag}\{V_1, V_2, \ldots, V_n\}$, $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)^\top$, $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_n)^\top$, and

$$w_i = \frac{(\mathrm{d}\mu_i/\mathrm{d}\eta_i)^2}{V_i}, \quad V_i = \mu_i[1 + W_0(\mu_i)], \quad i = 1, 2, \ldots, n.$$

Note that V_i is the variance function of Y_i. The Fisher information matrix for $\boldsymbol{\beta}$ is given by $\boldsymbol{K}(\boldsymbol{\beta}) = \boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{X}$.

The ML estimator $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_1, \widehat{\beta}_2, \ldots, \widehat{\beta}_p)^\top$ of $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^\top$ is obtained as the solution of $\boldsymbol{U}(\widehat{\boldsymbol{\beta}}) = \boldsymbol{0}_p$, where $\boldsymbol{0}_p$ denotes a p-dimensional vector of zeros. Unfortunately, the ML estimator $\widehat{\boldsymbol{\beta}}$ has no closed-form expression and hence its computation has to be performed numerically. One can use, for example, the Newton–Raphson iterative technique. On the other hand, the Fisher scoring method can be used to estimate $\boldsymbol{\beta}$ by iteratively solving the equation

$$\boldsymbol{\beta}^{(m+1)} = (\boldsymbol{X}^\top \boldsymbol{W}^{(m)} \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{W}^{(m)} \boldsymbol{z}^{(m)}, \tag{9}$$

where $m = 0, 1, \ldots$ is the iteration counter, $\boldsymbol{z} = (z_1, z_2, \ldots, z_n)^\top = \boldsymbol{\eta} + \boldsymbol{W}^{-1/2} \boldsymbol{V}^{-1/2} (\boldsymbol{y} - \boldsymbol{\mu})$ acts as a modified response variable in (9) whereas $\boldsymbol{W}$ is a weight matrix, and $\boldsymbol{\eta} = (\eta_1, \eta_2, \ldots, \eta_n)^\top$. Using Eq. (9) and any software with a weighted linear regression routine, the ML estimate $\widehat{\boldsymbol{\beta}}$ can be computed iteratively.

Under some mild regularity conditions and when n is large, it follows that $\widehat{\boldsymbol{\beta}} \overset{a}{\sim} \mathcal{N}_p(\boldsymbol{\beta}, \boldsymbol{K}(\boldsymbol{\beta})^{-1})$. Let β_r ($r = 1, 2, \ldots, p$) be rth component of $\boldsymbol{\beta}$. For $0 < \alpha < 1/2$, the asymptotic confidence interval for β_r is given by

$$\widehat{\beta}_r \pm \Phi^{-1}(1 - \alpha/2)\, \mathrm{se}(\widehat{\beta}_r), \quad r = 1, 2, \ldots, p.$$

The above interval has an asymptotic coverage of $100(1 - \alpha)\%$. Here, $\mathrm{se}(\widehat{\beta}_r)$ means the asymptotic standard error of $\widehat{\beta}_r$, which is obtained as the square root of the (r, r)th element of $\boldsymbol{K}(\widehat{\boldsymbol{\beta}})^{-1}$. Finally, $\Phi^{-1}(\cdot)$ denotes the standard normal quantile function.

The deviance of the Bell regression model is $D = \sum_{i=1}^{n} d_i^2(y_i, \widehat{\mu}_i)$, and the deviance components are

$$d_i^2(y_i, \widehat{\mu}_i) = 2 \begin{cases} \exp\left(1 - \mathrm{e}^{W_0(\widehat{\mu}_i)}\right), & y_i = 0, \\ \mathrm{e}^{W_0(\widehat{\mu}_i)} - \mathrm{e}^{W_0(y_i)} + y_i \log\left(\dfrac{W_0(y_i)}{W_0(\widehat{\mu}_i)}\right), & y_i > 0, \end{cases}$$

where $\widehat{\mu}_i = g^{-1}(\boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}})$ is the ML estimate of $\mu_i = g^{-1}(\boldsymbol{x}_i^\top \boldsymbol{\beta})$, for $i = 1, 2, \ldots, n$. Under some mild regularity conditions, it follows that $D \overset{a}{\sim} \chi^2_{n-p}$ and hence the deviance D may be used as a measure of goodness-of-fit for the Bell regression model fitted to real data; that is, the smaller the value of D, the better the fit to the real data.

Next, a small Monte Carlo simulation experiment is reported to explore the finite-sample behavior of the ML estimator of $\boldsymbol{\beta}$. The Monte Carlo experiments were carried out using

$$\log(\mu_i) = \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}, \quad i = 1, 2, \ldots, n, \tag{10}$$

where $x_{i1} = 1$ for $i = 1, 2, \ldots, n$, and the true values of the parameters were taken as $\beta_1 = 1.0$, $\beta_2 = 0.5$ and $\beta_3 = 1.5$. The values of x_{i2} were obtained as random draws of the standard normal distribution, and the values of x_{i3} were obtained

Table 1
Mean and SD of the MLEs, and CP of the 90% and 95% intervals for the regression parameters.

Parameter	$n = 50$				$n = 80$			
	Mean	SD	CP (90%)	CP (95%)	Mean	SD	CP (90%)	CP (95%)
β_1	0.986	0.213	0.905	0.955	0.990	0.166	0.901	0.950
β_2	0.504	0.086	0.897	0.951	0.501	0.079	0.901	0.951
β_3	1.500	0.342	0.900	0.952	1.503	0.248	0.904	0.952
Parameter	$n = 150$				$n = 300$			
	Mean	SD	CP (90%)	CP (95%)	Mean	SD	CP (90%)	CP (95%)
β_1	0.995	0.120	0.896	0.945	0.997	0.085	0.893	0.948
β_2	0.500	0.047	0.897	0.951	0.500	0.041	0.896	0.957
β_3	1.502	0.191	0.894	0.945	1.501	0.128	0.894	0.950

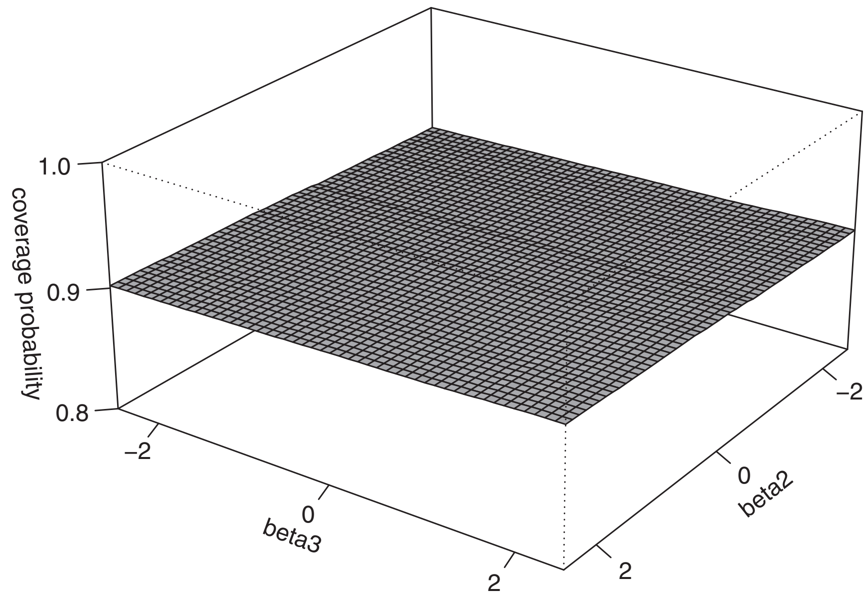

Fig. 2. CP of the 90% intervals for the regression parameters; $n = 80$.

as random draws of the uniform distribution on (0, 1). We consider $n = 50, 80, 150$ and 300. The covariate values were held constant throughout the simulations. We evaluate the point estimates by considering the following quantities: the mean and the standard deviation (SD). We also consider the coverage probability (CP) of the 90% and 95% intervals for the regression parameters. We use 5000 Monte Carlo replications. The numerical results are presented in Table 1. Note that the performance of the ML estimator of $\boldsymbol{\beta}$ is quite good, exhibiting small bias and respectable SD in all cases considered; that is, the ML estimates are quite stable and, more importantly, are very close to the true values. Additionally, as expected, the SDs decrease as the sample size increases. Note also that the empirical CPs are very close to the nominal levels in all cases.

We now set $\beta_1 = 1$ and $n = 80$ and evaluate the CP of the 90% interval for β_2 for a grid of values for (β_2, β_3). In Fig. 2 we report the results in a 3D plot. It is apparent from the plot that the CP is uniformly close to the nominal level in the range of parameters considered. Simulations for other settings yielded similar results.

Finally, we consider model (10) with the true values of the parameters taken as $\beta_1 = 1.0$, $\beta_2 = 1.2$ and $\beta_3 = 2.0$, and $n =$ 90 and 160. We consider three scenarios: (i) generating the data as $Y_i \sim \text{Poisson}(\mu_i)$ and estimating the parameters assuming that $Y_i \sim \text{Bell}(W(\mu_i))$; (ii) generating the data as $Y_i \sim \text{NB}(\mu_i, \kappa)$, with $\kappa = 10$, and estimating the parameters assuming that $Y_i \sim \text{Bell}(W_0(\mu_i))$; and (iii) generating the data as $Y_i \sim \text{Bell}(W_0(\mu_i))$ and estimating the parameters assuming that $Y_i \sim \text{NB}(\mu_i, \kappa)$. Here, $\text{NB}(\mu_i, \kappa)$ stands for the NB distribution with mean μ_i and dispersion parameter $\kappa > 0$. The NB distribution is usually employed for modeling count data that exhibit overdispersion. The variance of the $\text{NB}(\mu_i, \kappa)$ distribution is given by $\mu_i(1 + \mu_i/k)$, hence it is a second order polynomial in μ_i, while the variance of the Bell distribution is not far from linear for small values of the mean. These simulations are intended to shed some light on the effect of misspecification of the model assumed for the data. The numerical results are presented in Table 2. The figures in this table reveal that, despite the misspecification, the estimation is almost unbiased for all the scenarios. On the other hand, interval estimation is affected by misspecification. We noticed that the CP of the intervals for the regression parameters are very close to the nominal levels when the the model is correctly specified (results not shown). However, the CP under misspecified models may be

Table 2
Mean and SD of the MLEs, and CP of the 90% and 95% intervals for the regression parameters under misspecification.

Parameter	Scenario (i)							
	$n = 90$				$n = 160$			
	Mean	SD	CP (90%)	CP (95%)	Mean	SD	CP (90%)	CP (95%)
β_1	1.000	0.081	0.993	0.998	1.002	0.065	0.992	0.998
β_2	1.201	0.023	0.994	1.000	1.201	0.027	0.993	0.999
β_3	1.998	0.109	0.994	0.998	1.996	0.090	0.992	1.000
Parameter	Scenario (ii)							
	$n = 90$				$n = 160$			
	Mean	SD	CP (90%)	CP (95%)	Mean	SD	CP (90%)	CP (95%)
β_1	1.001	0.149	0.796	0.878	1.000	0.102	0.890	0.948
β_2	1.197	0.071	0.660	0.744	1.199	0.058	0.784	0.860
β_3	1.995	0.219	0.767	0.858	1.997	0.155	0.869	0.926
Parameter	Scenario (iii)							
	$n = 90$				$n = 160$			
	Mean	SD	CP (90%)	CP (95%)	Mean	SD	CP (90%)	CP (95%)
β_1	0.991	0.143	0.852	0.922	0.998	0.099	0.854	0.918
β_2	1.203	0.060	0.885	0.946	1.201	0.043	0.914	0.956
β_3	2.007	0.188	0.882	0.934	1.999	0.139	0.900	0.952
κ	9.971	4.402			8.574	2.772		

distorted. Our results suggest that diagnostic tools should be employed to check model adequacy in practical applications; see Section 5.

5. Residual analysis

It is well-known that residuals carry important information concerning the appropriateness of assumptions that underlie statistical models, and thereby play an important role in checking model adequacy. They are used to identify discrepancies between models and data, so it is natural to base residuals on the contributions made by individual observations to measures of model fit. Next, we introduce two residuals for the class of Bell regression models.

The first residual is given by

$$r_i^d = \mathrm{sign}(y_i - \widehat{\mu}_i)d_i(y_i, \widehat{\mu}_i), \quad i = 1, 2, \dots, n,$$

where $\widehat{\mu}_i = g^{-1}(\boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}})$ is the ML estimate of $\mu_i = g^{-1}(\boldsymbol{x}_i^\top \boldsymbol{\beta})$, and $d_i(y_i, \widehat{\mu}_i)$ was defined in Section 4. We shall call r_i^d as ith *deviance residual*. Let $\boldsymbol{H} = \boldsymbol{W}^{1/2}\boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{W}^{1/2}$ be the 'hat' matrix. The modified deviance residual is defined as

$$r_i^{md} = \frac{r_i^d}{\sqrt{1 - h_i}}, \quad i = 1, 2, \dots, n,$$

where $h_i = h_i(\widehat{\boldsymbol{\beta}})$ is the ith diagonal value of $\boldsymbol{H}$ evaluated at $\widehat{\boldsymbol{\beta}}$. Note that r_i^{md}, unlike r_i^d, takes the discrepant values in the covariate matrix $\boldsymbol{X}$ into account. The exact distribution of the above residuals is not known. As suggested by Atkinson [7, § 4.2], it is usual to add envelopes in the normal probability plots for these residuals (see, for instance, Neter et al. [8, § 14.6]) to decide whether the observed residuals are consistent with the fitted model.

For relatively large datasets, a visual tool may be employed to examine whether the mean-variance relationship implied by the Bell model is suitable for the data. Friendly [9, § 11.3.3] suggest to group the data according to the ordered fitted values of the linear predictor, compute the sample mean and variance for each group, plot the variances against the means along with a smoothed curve. The Bell variance curve and the variance curves of other competing models may be added for comparison.

6. Real data application

In this section, we shall consider the Bell regression model in a real data application for illustrative purposes. The R program (R Core Team [10]) was used to perform the computations, and the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm was used to estimate the model parameters. The R code to compute the estimates of the Bell regression model parameters is provided in the Appendix.

We are interested in modeling the number of faults (Y) in rolls of fabric of different lengths as a function of the length of the roll (x). We have 36 rolls of fabric of different lengths. The data are listed in Hinde [11]. The scatterplot of the data

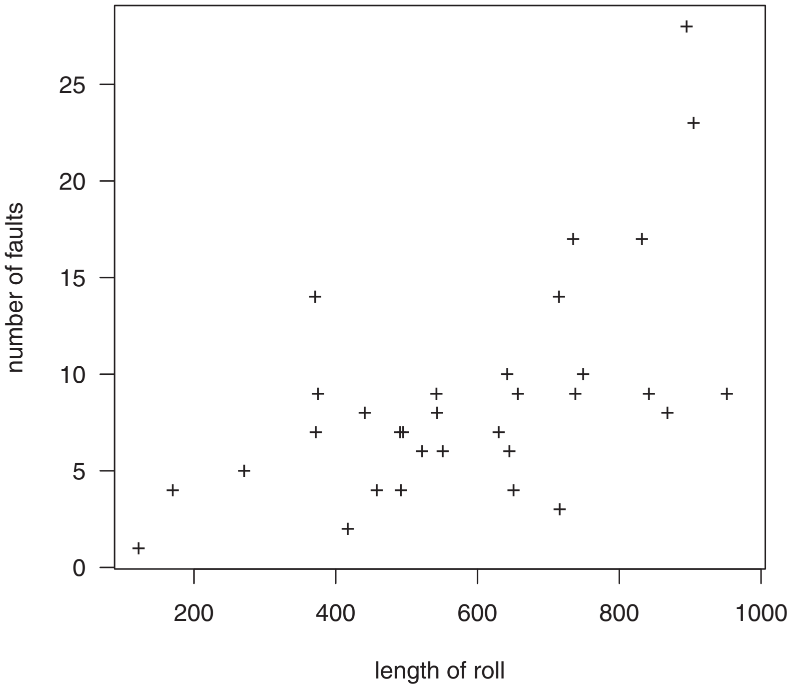

Fig. 3. Number of faults as a function of length of roll.

Table 3
Parameter estimates; Poisson regression.

Parameter	Estimate	SE	95% CI
β_1	0.9718	0.2125	(0.5553; 1.3883)
β_2	0.0019	0.0003	(0.0013; 0.0025)

Table 4
Parameter estimates; Bell regression.

Parameter	Estimate	SE	95% CI
β_1	0.9853	0.3337	(0.3312; 1.6394)
β_2	0.0019	0.0005	(0.0009; 0.0029)

is displayed in Fig. 3, which indicates that the mean and variance of the number of faults increase with the length of the roll. The first natural choice for modeling these data is the Poisson regression model. So, we initially assume that $Y_i \sim$ Poisson(μ_i), where

$$\log(\mu_i) = \eta_i = \beta_1 + \beta_2 x_i, \quad i = 1, 2, \ldots, 32.$$

Table 3 lists the ML estimates for the Poisson regression model parameters. We also provide the standard errors (SE) and the 95% confidence intervals (CI). The deviance is 61.758 on 30 degrees of freedom (p-value $= .0006$), which indicates overdispersion and a poor fit to the data. The deviance of the null (constant) model is 103.714 and hence the likelihood ratio (LR) statistic to test $H_0 : \beta_2 = 0$ is $\omega = 41.956$ (p-value $< .001$), indicating that the length of the roll is strongly significant. (The Wald test for the null hypothesis $H_0 : \beta_2 = 0$ also yields p-value $< .001$). Fig. 4 displays the modified deviance residuals and the corresponding normal probability plot (with simulated envelopes) for the Poisson model, which definitely discards this regression model to fit these data because many observations are outside the envelope.

Now, we assume that $Y_i \sim$ Bell($W_0(\mu_i)$), where

$$\log(\mu_i) = \eta_i = \beta_1 + \beta_2 x_i, \quad i = 1, 2, \ldots, 32.$$

Table 4 lists the ML estimates, asymptotic SEs and the 95% asymptotic CI for the Bell regression model parameters. The deviance is 23.175 on 30 degrees of freedom (p-value $= .808$), which indicates a good fit to the data. The deviance of the null (constant) model is 38.820 and hence the LR statistic for testing H_0: $\beta_2 = 0$ is $\omega = 15.645$ (p-value $< .001$), which reveals that the length of the roll is strongly significant. (The Wald test for the null hypothesis H_0: $\beta_2 = 0$ also delivers p-value $< .001$). Fig. 5 displays the modified deviance residuals and the corresponding normal probability plot (with simulated

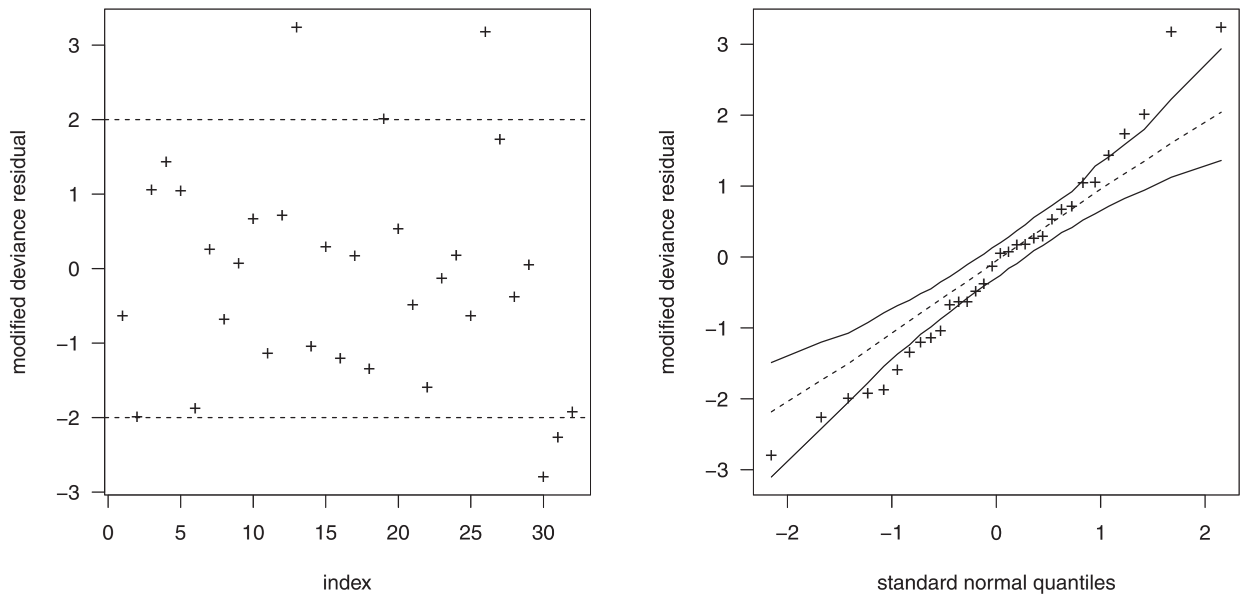

Fig. 4. Modified deviance residuals; Poisson regression model.

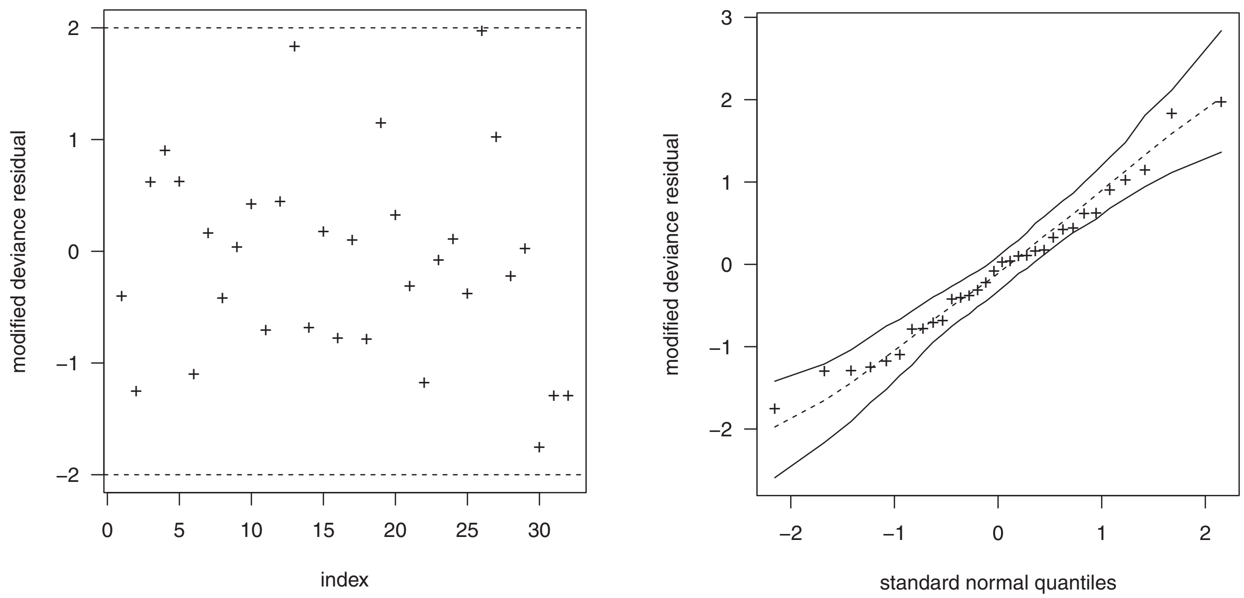

Fig. 5. Modified deviance residuals; Bell regression model.

envelopes) for the Bell model. From Fig. 5 we may conclude that the Bell regression model seems to be appropriate to fit these data, since no observation is outside the envelope; that is, the modified deviance residual plots do not present any unusual feature.

An interesting question is to verify how the familiar NB regression model fits these data. Undoubtedly, the NB regression model is the most popular model used in practice for modeling count data with overdispersion and hence there is a vast literature about this regression model. The reader is referred to the very important book by Hilbe [12]. So, we assume that $Y_i \sim$ NB(μ_i, κ), where

$$\log(\mu_i) = \eta_i = \beta_1 + \beta_2 x_i, \quad i = 1, 2, \ldots, 32,$$

and $\kappa > 0$ is the dispersion parameter, which is assumed to be unknown but the same for all observations. Table 5 lists the ML estimates, asymptotic SEs and the 95% asymptotic CI for the NB regression model parameters; the ML estimate of the dispersion parameter (SE between parentheses) is $\widehat{\kappa} = 9.57$ (4.89). The LR statistic for testing $H_0\colon \beta_2 = 0$ is $\omega = 16.508$ (p-value $< .001$), indicating that the length of the roll is strongly significant. (The Wald test for the null hypothesis $H_0\colon \beta_2 = 0$

Table 5
Parameter estimates; NB regression.

Parameter	Estimate	SE	95% CI
β_1	1.0015	0.2784	(0.4558; 1.5472)
β_2	0.0019	0.0004	(0.0011; 0.0027)

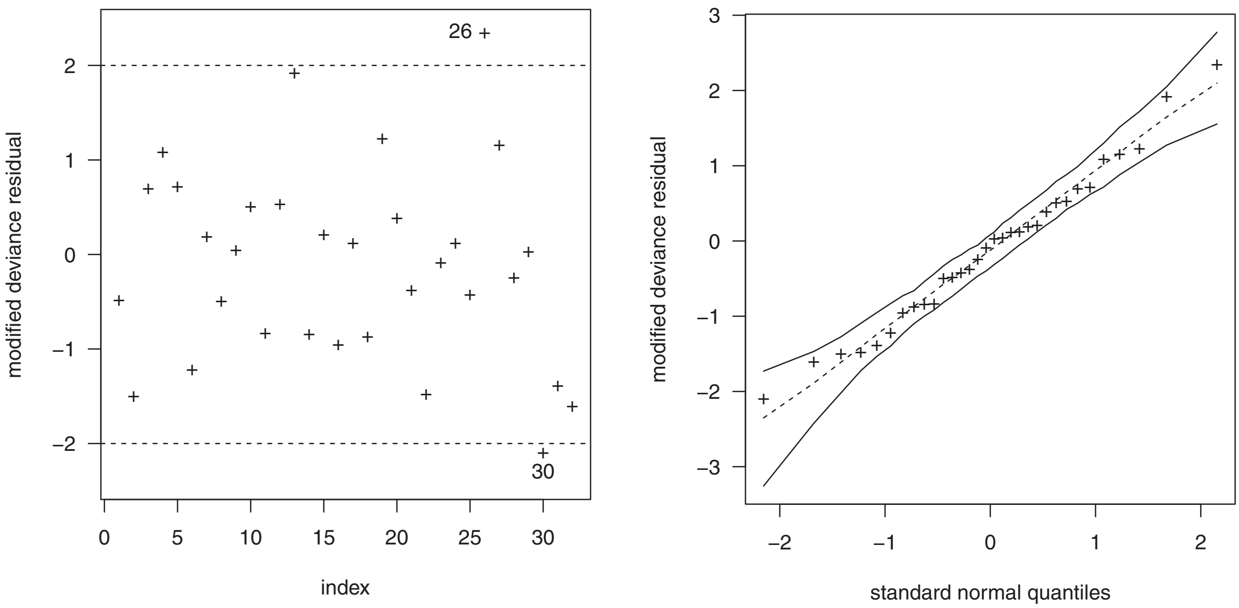

Fig. 6. Modified deviance residuals; NB regression model.

also yields p-value $< .001$). Fig. 6 displays the modified deviance residuals and the corresponding normal probability plot (with simulated envelopes) for the NB model, which indicates a large positive residual (case #26) and a large negative residual (case #30), unlike the Bell regression model (see Fig. 5). The case #26 corresponds to a roll which has length 371 with 14 faults, whereas the case #30 corresponds to a roll which has length 716 with only 3 faults. Similarly to the Bell regression model, the NB regression model is also appropriate to model these data, because no observation is outside the envelope (see Fig. 6).

Finally, a natural question is how to choose the best regression model between the Bell and NB models to fit the current data. The regression parameter estimates of these models are very similar, and the corresponding SEs of these overdispersion models, as would be expected, are larger than those for the Poisson model. Next, we shall consider the generalized LR statistic (V_{LR}) introduced in Vuong [13] to try to choose the best regression model. We have that $V_{\text{LR}} = \Lambda\Psi^{-1/2}$, with

$$\Lambda = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\log\left(\frac{\widehat{\mu}_{1i}}{\widehat{\mu}_{2i}}\right),$$

$$\Psi = \frac{1}{n}\sum_{i=1}^{n}\left[\log\left(\frac{\widehat{\mu}_{1i}}{\widehat{\mu}_{2i}}\right)\right]^2 - \left[\frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{\widehat{\mu}_{1i}}{\widehat{\mu}_{2i}}\right)\right]^2,$$

where $\widehat{\mu}_{1i} = \exp(0.9853 + 0.0019\,x_i)$ corresponds to the Bell regression model, whereas $\widehat{\mu}_{2i} = \exp(1.0015 + 0.0019\,x_i)$ corresponds to the NB regression model. Under the null hypothesis of equivalence of the models, the statistic V_{LR} has a standard normal distribution; see Vuong [13]. Let α be the significance level. If $|V_{\text{LR}}| \leq \Phi^{-1}(1-\alpha/2)$, then the null hypothesis of equivalence of the models is not rejected. On the other hand, we reject at significance level α the null hypothesis in favor of the Bell model being better (worse) than the NB model if $V_{\text{LR}} > \Phi^{-1}(1-\alpha)$ ($V_{\text{LR}} < -\Phi^{-1}(1-\alpha)$). The observed value of V_{LR} to test the equivalence of the Bell and NB models equals -0.576 (p-value $= .282$) and hence the test does not distinguish between the Bell and NB models to fit these data. It is noteworthy that the Poisson model is rejected in favor of the Bell model ($V_{\text{LR}} = 2.004$, and the corresponding p-value is .023) at significance levels as low as 2.5%. As remarked before, the Bell regression model does not need an additional parameter to accommodate overdispersion, unlike the NB regression. Therefore, at least in terms of parsimony, we may choose the Bell model as the best regression model to fit the data.

7. Concluding remarks

In this paper we introduce a new discrete distribution for count data, named as the *Bell distribution*. The new class of discrete distributions was obtained from a series expansion provided by Bell [1,2]. The new distribution is indexed by one parameter and it has a very simple form for its probability mass function. We have provided a comprehensive account of its structural properties, including explicit expressions for the probability generating function, moment generating function, characteristic function, etc. Additionally, on the basis of the Bell distribution, a new regression model where the response variable is a count was proposed. Employing the frequentist approach, the parameter estimation of the Bell regression model is conducted by maximum likelihood, and the Fisher information is derived. We also consider two residuals for the Bell regression model. Further, the methodology developed in this paper is illustrated by means of an empirical application. In conclusion, the Bell regression model may provide a rather flexible mechanism for fitting a wide spectrum of discrete real world data sets which may present overdispersion, and we hope that the new model may serve as an alternative model to the very familiar Poisson and NB models for modeling count data in several areas.

Acknowledgments

We thank the referees, the Associate Editor and the Editor for their comments and suggestions. We gratefully acknowledge the financial support of the Brazilian agencies FAPESP (grant 2012/21788–2), FAPEMIG and CNPq (grants 166448/2014-1, 304388/2014–9 and 301808/2016–3).

Appendix. The R code

```
## required package
require(LambertW)

## R function to estimate the Bell regression
## model parameters (link function = "log")
bell.reg <- function(formula, data)
{
    cl <- match.call()
    if (missing(data))
      data <- environment(formula)
    mf <- match.call(expand.dots = FALSE)
    m <- match(c("formula", "data"), names(mf), 0L)
    mf <- mf[c(1L, m)]
    mf$drop.unused.levels <- TRUE
    oformula <- as.formula(formula)
    mf$formula <- formula
    mf[[1L]] <- as.name("model.frame")
    mf <- eval(mf, parent.frame())
    mt <- terms(formula, data = data)
    Y <- model.response(mf, "numeric")
    X <- model.matrix(mf)
    if (length(Y) < 1)
      stop("empty model")
    if (!(min(Y) >= 0))
```

```
  stop("invalid dependent variable")
 floglikBell <- function(vPar){
   veta <- X%*%vPar
   vmu <- exp(veta)
   loglik <- sum(-exp(W(vmu)) + Y*log(W(vmu)))
   loglik
 }
 fscoreBell <- function(vPar){
   veta <- X%*%vPar
   vmu <- exp(veta)
   vt <- 1/(1 + W(vmu))
   mT <- diag(as.vector(vt))
   score <- t(X)%*%mT%*%(Y - vmu)
   score
 }
 fFisherBell <- function(vPar){
   veta <- X%*%vPar
   vmu <- exp(veta)
   vw <- vmu/(1 + W(vmu))
   mW <- diag(as.vector(vw))
   mF <- t(X)%*%mW%*%X
   mF
 }
    fit0 <- glm(formula, family=poisson(link=log))
    start <- fit0$coef
    opt <- optim(start, fn = floglikBell, gr = fscoreBell,
      method = "BFGS", control=list(fnscale=-1), hessian=FALSE)
    if (opt$conv != 0)
      stop("algorithm did not converge")
    beta <- opt$par
    se <- sqrt(diag(solve(fFisherBell(beta))))
    z.value <- beta/se
    p.value <- 1 - pnorm(abs(z.value))
    names(beta) <- colnames(X)
    rval <- cbind(round(beta, 6), round(se, 6),
                  round(z.value, 6), round(p.value, 6))
    colnames(rval) <- c("Estimate", "Std. Error",
                        "z value", "Pr(>|z|)")
    return(rval)
}

## Example: real data used in the paper
## nf = "number of faults" and lroll = "length of the roll"
nf <- c(6,4,17,9,14,8,5,7,7,7,6,8,28,4,10,4,
        8,9,23,9,6,1,9,4,9,14,17,10,7,3,9,2)
lroll <- c(551,651,832,375,715,868,271,630,491,
           372,645,441,895,458,642,492,543,842,
           905,542,522,122,657,170,738,371,735,
           749,495,716,952,417)
bell.fit <- bell.reg(nf ~ lroll)
bell.fit
            Estimate Std. Error  z value Pr(>|z|)
(Intercept) 0.985251   0.333664 2.952822 0.001574
lroll       0.001909   0.000492 3.878065 0.000053
```

References

[1] E.T. Bell, Exponential polynomials, Ann. Math. 35 (1934) 258–277.
[2] E.T. Bell, Exponential numbers, Am. Math. Mon. 41 (1934) 411–419.
[3] L. Jánossy, A. Rényi, J. Aczél, On composed poisson distributions: I, Acta Math. Acad. Sci. Hungaricae 1 (1950) 209–224.
[4] J. Touchard, Propriétés arihtmétiques de certains nombres récurrents, Ann. Soc. Sci. Bruxelles 53 (1933) 21–31.
[5] R.M. Corless, G.H. Gonnet, D.E.G. Hare, D. Jeffrey, D.E. Knuth, On the Lambert W function, Adv. Comput. Math. 5 (1996) 329–359.
[6] P. McCullagh, J. Nelder, Generalized Linear Models, second ed., Chapman & Hall, London, 1989.
[7] A.C. Atkinson, Plots, Transformation and Regression: An Introduction to Graphical Methods of Diagnostic Regression Analysis, Oxford University Press, New York, 1985.
[8] J. Neter, M.H. Kutner, C.J. Nachtsheim, W. Wasserman, Applied Linear Statistical Models, fourth ed., Irwin, Chicago, 1996.
[9] M. Friendly, D. Meyer, Discrete Data Analysis with R: Visualization and Modeling Techniques for Categorical and Count Data, CRC Press, Boca Raton, 2005.
[10] R.C. Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2016.
[11] J. Hinde, Compound poisson regression models, in: R. Gilchrist (Ed.), GLIM82, Springer, New York, 1982, pp. 109–121.
[12] J.M. Hilbe, Negative Binomial Regression, second ed., Cambridge University Press, New York, 2007.
[13] Q.H. Vuong, Likelihood ratio tests for model selection and non-nested hypotheses, Econometrica 57 (1989) 307–333.