A Simple and Useful Regression Model for Fitting Count Data

Rodrigo M. R. de Medeiros[1], Marcelo Bourguignon[2]
[1] Universidade de São Paulo
[2] Universidade Federal do Rio Grande do Norte

Abbreviated abstract: We introduce a varying dispersion regression model for count data where the response variable is BerG-distributed using a new parameterization of this distribution, which is indexed by mean and dispersion parameters. The model can fit count data when overdispersion, equidispersion, underdispersion, or zero inflation (or deflation) is indicated. The maximum likelihood method is used to estimate the model parameters. Also, we define hypothesis tests for the dispersion parameter and consider residual analysis. We conduct simulation studies and present empirical applications.

Related publications:

– Bourguignon, M., and de Medeiros, R.M.R, *TEST,* **31**, 790–827, (2022).
– Bourguignon, M., et all, *Statistical Papers,* **63**, 821–848, (2022).

4th Conference on
Statistics and
Data Science
Salvador, Brazil (online)
December 1-3, 2022

The BerG Distribution and its Associated Regression Model

I The BerG distribution

We say that a discrete random variable Y follows a BerG distribution with mean $\mu > 0$ and dispersion index $\phi > |\mu - 1|$, if its probability mass function is given by

$$\Pr(Y = y) = \frac{1 - \mu + \phi}{1 + \mu + \phi} I(y = 0) + 4\mu \frac{(\mu + \phi - 1)^{y-1}}{(\mu + \phi + 1)^{y+1}} I(y \in \mathbb{N}).$$

We write $Y \sim \mathrm{BerG}(\mu, \phi)$. In addition, the variance of Y is given by

$$\mathrm{Var}(Y) = \mu\phi.$$

II Regression structure

Let $Y_1, \dots, Y_n$ be a sample of n independent random variables, where

$$Y_i \sim \mathrm{BerG}(\mu_i, \phi_i), \qquad i = 1, \dots, n,$$

and we assume the following regression structure:

$$g_1(\mu_i) = \boldsymbol{x_i}^T \boldsymbol{\beta} \qquad \mathrm{e} \qquad g_2(\phi_i) = \boldsymbol{z_i}^T \boldsymbol{\gamma}.$$

Attractive features of the distribution

- Easy interpretation;
- The flexibility of modeling low counts;
- Mathematical simplicity.

Main contributions:

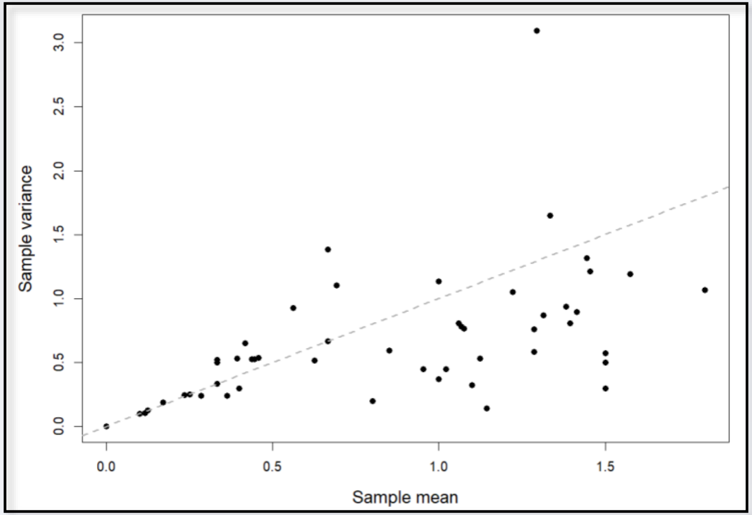
- A new parametrization;
- Modeling framework for count data:
  - Parameter estimation;
  - Hypothesis tests for constant dispersion;
  - Residual analysis;
- A new goodness-of-fit measure;
- An initial version of an R package: bergreg (available at GitHub).

Application: Number of Unharmed People in Road Traffic Accidents

Explanatory variables:

- **Weekend:** a two-level factor categorized into *weekday* and *weekend;*
- **Type:** a seven-level factor that identifies the type of accident, categorized into *frontal collision*, *side collision*, *transverse collision*, *rear-end collision*, *rollover*, *tipping*, and *other;*
- **Weather:** a four-level factor which identifies the weather conditions at the time of the accident, categorized as *clear sky*, *rain*, *cloudy*, and *other.*

Figure 1. Sample mean against sample variance of the response in each group formed by combinations of the levels of the explanatory variables.



4th Conference on Statistics and Data Science
Salvador, Brazil (online)
December 1-3, 2022

Results and Conclusion

Figure 2. Rootograms for the fitted regression models to the number of unharmed people by accident.

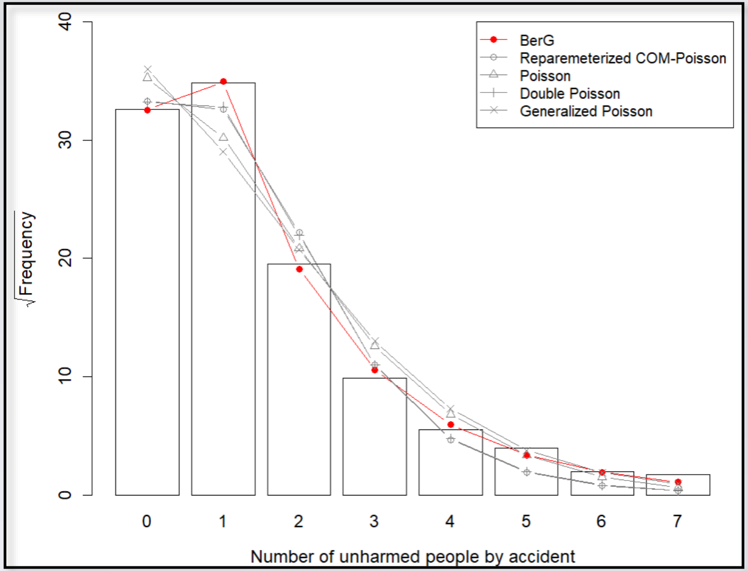


Table 1. Goodness-of-fit measures for the fitted regression models to the number of unharmed people by accident.

Model	AIC	BIC	HQIC	Pseudo- R^2
BerG	**5994.145**	**6124.846**	**6041.315**	**0.9997**
Rep. COM-Poisson	6074.776	6205.477	6121.946	0.9793
Poisson	6377.281	6442.632	6400.866	0.9264
Double Poisson	6066.828	6197.529	6113.998	0.9832
Generalized Poisson	6332.891	6463.591	6380.060	0.8863

4th Conference on
Statistics and Data Science
Salvador, Brazil (online)
December 1-3, 2022