The Journal of Open Source Software

fitODBOD: An R Package to Model Binomial Outcome Data using Binomial Mixture and Alternate Binomial Distributions.

Amalan Mahendran[1] **and Pushpakanthie Wijekoon**[1]

1 Department of Statistics and Computer Science, Faculty of Science, University of Peradeniya.

DOI: 10.21105/joss.01505

Software

- Review
- Repository
- Archive

Submitted: 13 June 2019
Published: 02 July 2019

License
Authors of papers retain copyright and release the work under a Creative Commons Attribution 4.0 International License (CC-BY).

Summary

The **R** package `fitODBOD` can be used to identify the best-fitting model for Over-dispersed Binomial Outcome Data (BOD). The Triangular Binomial (TriBin), Beta-Binomial (BetaBin), Kumaraswamy Binomial (KumBin), Gaussian Hypergeometric Generalized Beta-Binomial (GHGBB), Gamma Binomial (GammaBin), Grassia II Binomial (GrassiaIIBin) and McDonald Generalized Beta-Binomial (McGBB) distributions in the Family of Binomial Mixture Distributions (FBMD) are considered for model fitting in this package. Alternate Binomial Distributions such as Additive Binomial (AddBin), Beta-Correlated Binomial (BetaCorrBin), COM Poisson Binomial (COMPBin), Correlated Binomial (CorrBin), Lovinson Multiplicative Binomial (LMBin) and Multiplicative Binomial (MultiBin) distributions are used as well, replacing the traditional binomial distribution. Further, Probability Mass Function (PMF), Cumulative Probability Mass Function (CPMF), Negative Log Likelihood, Over-dispersion and parameter estimation (shape and distribution distinct parameters) can be explored for each fitted model with the `fitODBOD` package.

Introduction

Statistical methods are widely used for research in most disciplines. There is a focus towards fitting distributions to given data since the distributions of data depends on the method of data collection. For example, consider a binomial experiment where a fair coin is being tossed n times. Let the event of landing heads-up be defined as the success of probability p. Then, the number of heads out of n tosses is considered to be a single binomial variable, Y. Also if similar binomial experiments occur in N different clusters, a collection of $Y1$, $Y2$, $Y3$, ..., YN would form the BOD. Such data are frequently mentioned in fields of toxicology, biology, clinical medicine, epidemiology and many more. One may attempt to fit the BOD using the traditional binomial distribution, as it is characterized using the number of identical trials n and the probability of success parameter p. The parameter p ($p \in [0, 1]$) is usually assumed to be a constant from trial to trial and the trials are independent. In many empirical situations, it has been frequently observed that the actual observed variance of the BOD is greater than the assumed theoretical binomial variance. This outcome is typically known as "over-dispersion" (Anderson, 1988; Cox, 1983). Over-dispersion in BOD can occur either with a probability of success parameter p varying from trial to trial or if there is a correlation among binary trials. However, Collett (1991) argued that the above two cases of over-dispersion are frequently the same.

New distributions emerged to fit the BOD replacing the traditional binomial distribution. Li, Huang, & Zhao (2011) have developed the Kumaraswamy Binomial distribution, Rodriguez-Avi, Conde-Sanchez, Saez-Castillo, & Olmo-Jiminez (2007) have constructed the Gaussian

Hypergeometric Generalized Beta-Binomial distribution, Karlis & Xekalaki (2008) wrote the article on the Triangular Binomial distribution. Also, Grassia (1977) mentioned the Gamma Binomial and Grassia II Binomial distributions. The Beta-Binomial distribution is clearly explained in Johnson, Kotz, & Balakrishnan (1995). Initially the concept of mixing the binomial distribution with a unit bounded continuous distribution was done by Horsnell (1957), which led to the Uniform Binomial distribution. Recently, Manoj, Wijekoon, & Yapa (2013) had developed the McDonald Generalized Beta-Binomial distribution. Based on this research only the `fitODBOD` (version 1.1.0) package was released to CRAN in February, 2018. Recently this package became available on GitHub and has its own website, which has made the package more convenient for researchers who intend to use it.

Further, new types of binomial distributions were developed replacing the traditional binomial distribution, which are called Alternate Binomial Distributions. Paul (1985) has developed the Multiplicative Binomial distribution, while recently Elamir (2013) has done more research to form the Lovinson Multiplicative Binomial distribution. COM Poisson Binomial distribution was introduced first by Borges, Rodrigues, Balakrishnan, & Bazán (2014). The comparison of Beta-Correlated Binomial distribution with Correlated Binomial distribution was done by Kupper & Haseman (1978). Version 1.4.1 of `fitODBOD` (Mahendran & Wijekoon (2019)) holds all the distributions mentioned above and in the future more distributions developed to fit the BOD will be added to the package as major version updates.

Modelling

To fit a Binomial Mixture distribution for a raw BOD set, the following steps have to be used when using this package.

1. Extract the data in a meaningful way (**BODextract** function).
2. Check whether the binomial distribution can be fitted and if not test the over-dispersion (**fitBin** function) by using Pearson Chi-square goodness of fit test.
3. If over-dispersion exists, estimate the parameters for each distribution for the given data separately (**EstMLETriBin**, **EstMLEBetaBin**, **EstMGFBetaBin**, **EstMLEKumBin**, **EstMLEGammaBin**, **EstMLEGrassiaIIBin**, **EstMLEGHGBB**, **EstMLEMcGBB**, **EstMLEAddBin**, **EstMLEBetaCorrBin**, **EstMLECOMPBin**, **EstMLECorrBin**, **EstMLELMBin**, **EstMLEMultiBin** functions).
4. Based on the above estimated parameters corresponding models can be fitted (**fitTriBin**, **fitBetaBin**, **fitKumBin**, **fitGammaBin**, **fitGrassiaIIBin**,**fitGHGBB**, **fitMcGBB**, **fitAddBin**, **fitBetaCorrBin**, **fitCOMPBin**, **fitCorrBin**, **fitLMBin**, **fitMultiBin** functions).
5. Finally, compare the results and choose the best-fitting distribution for the data by using a plot or table.

Series of code to complete the steps from 1 to 5 are thoroughly discussed in the README file in the GitHub repository.

Conclusion

The `fitODBOD` package is constructed for the main purpose of fitting the given BOD and being able to choose the best-fitted Binomial Mixture and/or Alternate Binomial Distributions. The package has functions to calculate PMF, CPMF and Negative Log Likelihood of Triangular Binomial, Beta-Binomial, Kumaraswamy Binomial, Gamma Binomial, Grassia II Binomial, GHGBB, McGBB, Additive Binomial, Beta-Correlated Binomial, COM Poisson Binomial, Correlated Binomial, Lovinson Multiplicative and Multiplicative Binomial distributions. Further,

there are functions for probability density, cumulative density and moment about zero values for Triangular, Beta, Kumaraswamy, Gamma, Gaussian Hypergeometric Generalized Beta and Generalized Beta of First kind distributions. Using the steps outlined above, the best-fitting Binomial Mixture Distribution and/or Alternate Binomial Distribution is determined.

Main Dependencies

`fitODBOD` package has three main dependencies from CRAN. Functions from `hypergeo` are used for applications of GHGBB and Gaussian Hypergeometric Generalized Beta distribution. `stats` functions are used for integration situations for the Triangular Binomial distribution. Finally, `bbmle` package is used for the parameter estimation of ABD and FBMD under the concept of Maximum Likelihood Estimation.

References

Anderson, D. A. (1988). Some models for Overdispersed Binomial Data. *Australian & New Zealand Journal of Statistics*, *30*(2), 125–148. doi:10.1111/j.1467-842X.1988.tb00844.x

Borges, P., Rodrigues, J., Balakrishnan, N., & Bazán, J. (2014). A COM-Poisson type generalization of the binomial distribution and its properties and applications. *Statistics & Probability Letters*, *87*, 158–166. doi:10.1016/j.spl.2014.01.019

Collett, D. (1991). *Modelling Binary Data*. Chapman & Hall/CRC Texts in Statistical Science. Chapman & Hall. doi:10.1002/pst.100

Cox, D. R. (1983). Some Remarks on Overdispersion. *Biometrika*, *70*(1), 269–274. doi:10.1093/biomet/70.1.269

Elamir, E. A. (2013). Multiplicative-Binomial Distribution: Some Results on Characterization, Inference and Random Data Generation. *Journal of Statistical Theory and Applications*, *12*(1), 92–105. doi:10.2991/jsta.2013.12.1.8

Grassia, A. (1977). On a Family of Distributions with argument between 0 and 1 obtained by transformation of the Gamma and Derived Compund Distributions. *Australian Journal of Statistics*, *19*(2), 108–114. doi:10.1111/j.1467-842X.1977.tb01277.x

Horsnell, G. (1957). Economical Acceptance Sampling Schemes. *Journal of the Royal Statistical Society. Series A (General)*, *120*(2), 148–201. doi:10.2307/2342822

Johnson, N., Kotz, S., & Balakrishnan, N. (1995). *Continuous Univariate Distributions*. Wiley series in probability and mathematical statistics: Applied probability and statistics. Wiley & Sons.

Karlis, D., & Xekalaki, E. (2008). The Polygonal Distribution. Advances in mathematical and statistical modeling, 21–33. doi:10.1007/978-0-8176-4626-4_2

Kupper, L. L., & Haseman, J. K. (1978). The Use of a Correlated Binomial Model for the Analysis of Certain Toxicological Experiments. *Biometrics*, *34*(1), 69–76. doi:10.2307/2529589

Li, X., Huang, Y., & Zhao, X. (2011). The Kumaraswamy Binomial Distribution. *Chinese Journal of Applied Probability and Statistics*, *27*(5), 511–521.

Mahendran, A., & Wijekoon, P. (2019). *fitODBOD: Modeling Over Dispersed Binomial Outcome Data Using BMD and ABD*. Retrieved from https://CRAN.R-project.org/package=fitODBOD

Manoj, C., Wijekoon, P., & Yapa, R. D. (2013). The McDonald Generalized Beta-Binomial Distribution: A New Binomial Mixture Distribution and Simulation Based Comparison with Its Nested Distributions in Handling Overdispersion. *International Journal of Statistics and Probability*, *2*(2), 24–41. doi:10.5539/ijsp.v2n2p24

Paul, S. (1985). A three-parameter generalization of the Binomial Distribution. *Communications in Statistics - Theory and Methods*, *14*(6), 1497–1506. doi:10.1080/03610928508828990

Rodriguez-Avi, J., Conde-Sanchez, A., Saez-Castillo, A. J., & Olmo-Jiminez, M. J. (2007). A Generalization of the Beta-binomial Distribution. *Journal of the Royal Statistical Society. Series C: Applied Statistics*, *56*(1), 51–61. doi:10.1111/j.1467-9876.2007.00564.x