Accident Analysis and Prevention 107 (2017) 186–194

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

A methodology to design heuristics for model selection based on the characteristics of data: Application to investigate when the Negative Binomial Lindley (NB-L) is preferred over the Negative Binomial (NB)

Mohammadali Shirazi[a,*], Soma Sekhar Dhavala[b], Dominique Lord[a], Srinivas Reddy Geedipally[c]

[a] Zachry Department of Civil Engineering, Texas A & M University, College Station, TX 77843, United States
[b] Perceptron Learning Solutions Pvt Ltd, Bengaluru, India
[c] Texas A & M Transportation Institute, Arlington, TX 76013, United States

ARTICLE INFO

Keywords:
Model Selection
Heuristics
Characteristics of Data
Machine Learning
Negative Binomial
Negative Binomial Lindley

ABSTRACT

Safety analysts usually use post-modeling methods, such as the Goodness-of-Fit statistics or the Likelihood Ratio Test, to decide between two or more competitive distributions or models. Such metrics require all competitive distributions to be fitted to the data before any comparisons can be accomplished. Given the continuous growth in introducing new statistical distributions, choosing the best one using such post-modeling methods is not a trivial task, in addition to all theoretical or numerical issues the analyst may face during the analysis. Furthermore, and most importantly, these measures or tests do not provide any intuitions into why a specific distribution (or model) is preferred over another (Goodness-of-Logic). This paper ponders into these issues by proposing a methodology to design heuristics for Model Selection based on the *characteristics* of data, in terms of descriptive summary statistics, *before* fitting the models. The proposed methodology employs two analytic tools: (1) Monte-Carlo Simulations and (2) Machine Learning Classifiers, to design easy heuristics to predict the label of the 'most-likely-true' distribution for analyzing data. The proposed methodology was applied to investigate when the recently introduced Negative Binomial Lindley (NB-L) distribution is preferred over the Negative Binomial (NB) distribution. Heuristics were designed to select the 'most-likely-true' distribution between these two distributions, given a set of prescribed summary statistics of data. The proposed heuristics were successfully compared against classical tests for several real or observed datasets. Not only they are easy to use and do not need any post-modeling inputs, but also, using these heuristics, the analyst can attain useful information about why the NB-L is preferred over the NB - or vice versa- when modeling data.

1. Introduction

There has been a phenomenal growth in introducing novel distributions and models to analyze crash data over the last decade (see Lord and Mannering, 2010; Mannering and Bhat, 2014). Selecting the most appropriate and logically sound sampling distribution among all these alternatives plays a crucial role in modeling and further systematic safety analyses or evaluations, and has always been a subject of interest to safety scientists or researchers. So far, the comparison of distributions (or models) has usually been accomplished during the post-modeling phase − once data are fitted to all competitive alternatives, using measures such as the Goodness-of-Fit (GoF) statistics or the Likelihood Ratio Test (LRT). However, such metrics are neither easy to compute nor practically doable on some instances when many alternatives exist and/or when the analyst deals with big data. In addition, and most importantly, these metrics do not provide any intuitions into why one distribution is preferred over another or the logic behind the Model Selection (Goodness-of-Logic, as illustrated by Miaou and Lord, 2003). In this research, we address these topics, and contribute to the crash data modeling by introducing a methodology that provides heuristics to select the 'most-likely-true' sampling distribution among its competitors, based on *characteristics* of data, reflected into certain summary statistics, *before* fitting the competitive models based on their distributions.

The research in this study was motivated first by looking at the characteristics of the Poisson and Negative Binomial (NB) distributions. The analyst can choose between the Poisson and NB distributions just by looking at the mean (μ) and variance (σ^2) of the data, before fitting the distributions or models. A general rule of thumb is that, when data show a sign of over dispersion (i.e., when $\sigma^2/\mu > 1$), the analyst can

* Corresponding author.
E-mail addresses: alishirazi@tamu.edu (M. Shirazi), soma.dhavala@gmail.com (S.S. Dhavala), d-lord@tamu.edu (D. Lord), srinivas-g@tti.tamu.edu (S.R. Geedipally).

http://dx.doi.org/10.1016/j.aap.2017.07.002
Received 23 February 2017; Received in revised form 25 May 2017; Accepted 4 July 2017
Available online 05 September 2017
0001-4575/

move from 'Poisson' to 'NB'. In this case, the variance-to-mean-ratio (VMR) serves as a heuristic for Model Selection and the VMR greater than one as a "switching" point. Second, the research problem can be motivated by looking at the characteristics of the NB and Negative Binomial Lindley (NB-L) (Zamani and Ismail, 2010; Lord and Geedipally, 2011; Geedipally et al., 2012) distributions. Both of these distributions can handle over dispersion; however, the NB-L distribution is preferred when data are characterized by many zeros and/or have a heavy (or long) tail (Lord and Geedipally, 2011). Although we know the NB-L distribution performs better when data are skewed, it is not clear at what 'point' the analyst should shift from the 'NB' to the 'NB-L'. In other words, it is not explicitly clear, for example, what the skewness of data should be to prefer the NB-L distribution over the simple NB distribution. Is skewness the only measure to look at while deciding so? We develop a systematic approach to answer such questions.

The comparison between the NB and NB-L distributions is used as a case study to illustrate the proposed methodology. As discussed in Lord and Mannering (2010), the NB distribution, despite all its limitations, still remains the most common sampling distribution used by safety modelers or practitioners, due to its simplicity. The analyst, however, should be cautious about the NB shortcomings when modeling crash data. As such, the NB distribution does not perform well when data are characterized by excess number of zero responses or have a long (or heavy) tail. The NB-L distribution attempts to overcome such issues by mixing the NB with the Lindley distribution. Although the NB-L, or other advanced distributions, may have a better performance than the NB, they come at a cost of a more complicated modeling and consuming more computational resources. In practice, it can be argued that the NB distribution should generally be good until a certain point, at which we may need to switch to a better but more complex distribution, such as the NB-L. An important question should now be asked: At what point should a more complex distribution such as the NB-L be used instead of the NB? Model Selection heuristics will be proposed to address this question.

The idea of the paper can now be introduced: what are the "switching" points to move from one distribution to another when two or more competitive distributions are available? Can we predict the model to be used based on *characteristics* of the data, reflected in its summary statistics, to find the 'most-likely-true' sampling distribution *before* fitting the model? The objectives of this study consequently are: (1) document a methodology to design heuristics to decide between two or more competitive distributions, based on summary statistics of data; (2) apply the methodology to investigate the "switching" points (or heuristics, to be exact) to select the 'most-likely-true' distribution between the NB and NB-L distributions to model crash or other safety related data.

2. Methodology

At the heart of the proposed methodology lies a paradigm shift in how Model Selection is both viewed and treated. We view Model Selection as a classification problem − that is, given a set of discriminating features of the data, we like to predict the model that must have produced the observed data. It becomes a binary classification problem when the number of alternatives is two. This way of looking at Model Selection as a classification problem was first introduced, according to the authors' knowledge, by Pudlo et al. (2015), in the context of Approximate Bayesian Computation. Learning the both discriminating function and its arguments have traditionally been based on GoF or other Model Selection criteria such as the LRT, Akaike Information Criteria (AIC) and the likes. The discriminating function in such methods, which favor one model to the other, is often a simple comparator. A benefit of viewing the Model Selection as a classification problem is that we can take computational approach to learning a complex discriminating function based on simple descriptive statistics of the data.

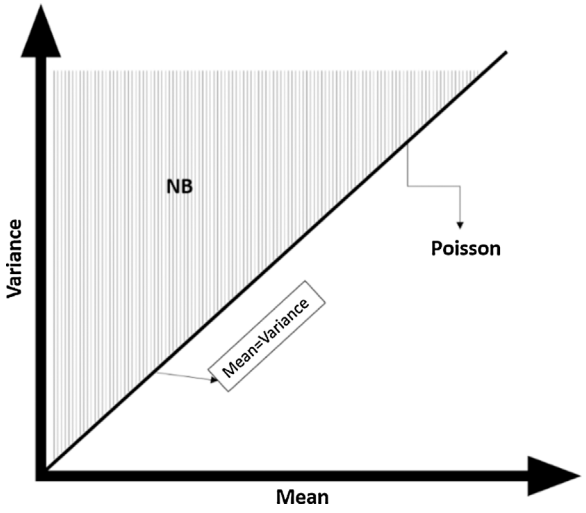

Fig. 1. Classifying the NB and Poisson Distributions Based on the Mean and Variance of the Population.

To clarify the strategy, let us assume the analyst is interested in choosing between the Poisson and NB distributions, based on the population 'mean' and 'variance'. We like to come up with a function that maps these two statistics to a label: '0′ for Poisson and '1′ for NB. The choice of the labels is completely arbitrary. The 'mean' and 'variance' of population would create a two dimensional (a flat plane) predictor space (Ω) for making decisions. Now, the analyst's task is to partition the predictor space and assign a label to each partition. We know that if the population VMR is greater than one (VMR $>$ 1), we may choose the NB distribution and if it is equal to one (VMR = 1), the Poisson distribution will be the preferred sampling distribution to use. Hence, the predictor space (Ω) can be classified between the Poisson and NB distributions in a way that is shown in Fig. 1.

The decision based on the VMR statistic, in this case, serves as a heuristic to select the 'most-likely-true' sampling distribution between the Poisson and NB distributions. It does not require fitting the models, estimating the model parameters, computing the test statistics, etc. It simply uses the descriptive statistics to arrive at a model recommendation.[1] When working with data, the 'population' VMR essentially is replaced with its 'sample' counterpart ($\widehat{\text{VMR}}$) and the decision based on observed data will be essentially the analyst best guess. Like any Model-Selection decisions, there is a chance that the decision based on a sample version of the VMR may be incorrect; this uncertainty can be quantified in terms of standard classifier performance metrics, such as false-positive-rate, Area under the Curve (AUC), and many others (Hastie et al., 2001; James et al., 2013).

In the case of 'Poisson' vs. 'NB', we knew, theoretically, how the two-dimensional predictor space should be partitioned between the Poisson and NB distributions; however, what if such insight was not available to us? In the absence of readily available analytical insights to guide Model Selection, we resort to computational approaches. It will be assumed that the distributions under consideration can be classified by 'm' summary statistics. These summary statistics would create an 'm-dimensional' predictor space; then, the analyst can benefit from two analytic tools, (1) Monte-Carlo Simulations, and (2) Machine Learning Classifiers, to partition the assumed m-dimensional predictor space between the competitive distributions.

Let us assume $\{A_1, A_2, \ldots, A_r\}$ and $\{S_1, S_2, \ldots, S_m\}$, respectively, denote a set of 'r' competitive distributions and 'm' types of summary statistics. We need to partition the m-dimensional predictor space that

[1] In Section 4, we show that there are strong correlations between the decision based on the VMR heuristic and the LRT statistic.