Computational Statistics and Data Analysis 107 (2017) 162–175

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Saddlepoint tests for accurate and robust inference on overdispersed count data

William H. Aeberhard[a,b,*], Eva Cantoni[a], Stephane Heritier[c]

[a] Research Center for Statistics and Geneva School of Economics and Management, University of Geneva, 1211 Geneva, Switzerland

[b] Sydney School of Public Health, University of Sydney, Sydney, NSW 2006, Australia

[c] Department of Epidemiology and Preventive Medicine, Monash University, Melbourne, VIC 3004, Australia

ARTICLE INFO

Article history:
Received 19 February 2016
Received in revised form 3 October 2016
Accepted 10 October 2016
Available online 25 October 2016

Keywords:
Exponential tilting
Negative binomial regression
Robust bounded influence tests
Small samples

ABSTRACT

Inference on regression coefficients when the response variable consists of overdispersed counts is traditionally based on Wald, score and likelihood ratio tests. As the accuracy of the p-values of such tests becomes questionable in small samples, three recently developed tests are adapted to the negative binomial regression model. The non-trivial computational aspects involved in their implementation, some of which remained obscure in the literature until now, are detailed for general M-estimators. Under regularity conditions, these tests feature a relative error property with respect to the asymptotic chi-squared distribution, thus yielding highly accurate p-values even in small samples. Extensive simulations show how these new tests outperform the traditional ones in terms of actual level with comparable power. Moreover, inference based on robust (bounded influence) versions of these tests remains reliable when the sample does not entirely conform to the model assumptions. The use of these procedures is illustrated with data coming from a recent randomized controlled trial, with a sample size of 52 observations. An R package implementing all tests is readily available.

1. Introduction

Even at the age of big data, drawing accurate inference in small samples outside the Gaussian world remains a challenging problem. When modeling overdispersed count data, where the variance of the data generating process exceeds its expectation (see Hinde and Demétrio, 1998, for an overview), inference is traditionally based on tests such as the Wald, score (Lagrange multiplier) and likelihood ratio tests. In this context, the convergence of these traditional test statistics to the χ^2 distribution under a null hypothesis H_0 relies on (first-order) asymptotic results. This implies that when using the χ^2 distribution as an approximation to the true (finite sample size) distribution of such test statistics, the approximation error is absolute and generally of order $O(n^{-1/2})$, where n is the sample size. This approximation error may be improved, for instance by correcting traditional test statistics with Bartlett factors (see e.g. Cordeiro, 1983; DiCiccio et al., 1991) or by resorting to resampling techniques such as the iterated bootstrap (Beran, 1987; Hall and Martin, 1988). Such means typically reduce the order of the approximation error to $O(n^{-3/2})$ (e.g. Cordeiro and Ferrari, 1991), but still in absolute terms and often at a substantial computational cost.

* Correspondence to: Department of Mathematics and Statistics, Dalhousie University, Halifax NS, Canada B3H 4R2.
E-mail addresses: william.aeberhard@dal.ca (W.H. Aeberhard), eva.cantoni@unige.ch (E. Cantoni), stephane.heritier@monash.edu (S. Heritier).

http://dx.doi.org/10.1016/j.csda.2016.10.009
0167-9473/

Relative errors, as opposed to absolute errors, represent greater precision and are in fact much more relevant when assessing the accuracy of approximations of small quantities such as p-values. The so-called saddlepoint test statistic introduced by Robinson et al. (2003) features a relative error of order $O(n^{-1})$ when approximating its sampling distribution with a χ^2 and thus leads to highly accurate inference even in small samples. Lô and Ronchetti (2009) have adapted this test to the generalized linear model (GLM) with canonical link and illustrated its excellent finite sample properties when based on maximum likelihood (ML) and robust bounded influence estimators (Hampel et al., 1986). For more general settings, an empirical (non-parametric) version of the saddlepoint test is available, and has been extended to different cases (e.g. Czellar and Ronchetti, 2010; Ma and Ronchetti, 2011; Lunardon and Ronchetti, 2014), as well as a recent alternative proposed by Camponovo and Otsu (2015), called the tilted exponential tilting test. A common feature of these three new tests is that they replace the computational effort of bootstrap-related techniques by numerical optimization. Although these three tests are asymptotically equivalent, their finite-sample performances have never been compared. Furthermore, some aspects of their intricate computation remain obscure in the literature, potentially discouraging statisticians to implement them.

ML based estimation and inference are well-known to be non-robust, in the sense that even slight departures from the model assumptions can drive estimates and test statistics to arbitrary values. Robust (bounded influence) tests (Heritier and Ronchetti, 1994) by contrast maintain their level and power under contamination. When concerned with the accuracy of p-values, the need for robust tests is all the more so important since the inaccuracies induced by deviations from the model likely dominate any precision gains coming from the use of higher-order asymptotics; see Ronchetti and Ventura (2001).

In this paper, we extend the three aforementioned tests to the negative binomial (NB) regression model, which lies outside the framework of Lô and Ronchetti (2009). This model is arguably one of the most popular models for overdispersed count data, and is a typical choice for the modeling of falls in clinical trials as in our motivating example in Section 5. This focus on the NB model does not preclude our notation, based on M-estimators, from being fairly general. Our exposition focuses on the computational aspects of the implementation of these three tests, which certainly go beyond the NB model. We propose both classical and robust versions of each test, the latter following Aeberhard et al. (2014). Through a large-scale simulation study, we investigate on the one hand how well, compared to traditional tests, the three new tests maintain their nominal level under H_0 and on the other hand how they perform in terms of power under local alternatives. Generating data both according to the assumed NB model and under contamination allows us to highlight the gain in inference stability of robust tests. To our knowledge, no such finite-sample comparison of these tests has been made so far, especially regarding power. R code (R Core Team, 2015) is provided as Supplementary Material (see Appendix A) for illustrating the use of our fully-documented R package which implements all tests studied here.

This paper is organized as follows: Section 2 sets our notation; in Section 3, we detail both theoretical and empirical versions of the saddlepoint test and the tilted exponential tilting test; Section 4 shows the results of our extensive Monte Carlo simulation study; Section 5 presents the application of these tests to falls data coming from a recent randomized controlled trial; finally, Section 6 concludes this paper.

2. Framework

Our sample consists of independent scalar count responses Y_i coupled with a vector of covariates $\boldsymbol{x}_i \in \mathbb{R}^p$, for $i = 1, \dots, n$. We assume $Y_i|\boldsymbol{x}_i$ follows a NB distribution with probability mass function (pmf)

$$\Pr(Y_i = y; \sigma, \mu_i) = \frac{\Gamma(y + 1/\sigma)}{\Gamma(1/\sigma)\Gamma(y+1)} (\sigma\mu_i + 1)^{-1/\sigma} \left(\frac{\sigma\mu_i}{\sigma\mu_i + 1}\right)^y, \tag{1}$$

where the mean parameter $\mu_i = \mathrm{E}[Y_i|\boldsymbol{x}_i]$ is linked to the linear combination of covariates through a natural logarithm, $\mu_i = \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta})$ with $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_{p-1})^\top \in \mathbb{R}^p$, and the overdispersion parameter $\sigma > 0$ appears in the quadratic variance function $\mathrm{Var}[Y_i|\boldsymbol{x}_i] = V(\mu_i) = \mu_i + \sigma\mu_i^2$. The Poisson GLM with mean μ_i is a limiting case when $\sigma \to 0$.

Let $\boldsymbol{z}_i = (Y_i, \boldsymbol{x}_i^\top)^\top$ and $\boldsymbol{\theta} = (\sigma, \boldsymbol{\beta}^\top)^\top$. We want to make inference about a subset of $\boldsymbol{\beta}$ of dimension $d < p$ and thus consider σ and the first $(p - d)$ elements of $\boldsymbol{\beta}$ as nuisance parameters. Let $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^\top, \boldsymbol{\theta}_2^\top)^\top$, so that the parameter of interest is $\boldsymbol{\theta}_2 \in \mathbb{R}^d$. This leads to testing the composite hypothesis $H_0 : \boldsymbol{\theta}_2 = \boldsymbol{\theta}_{2,0}$ against a two-sided alternative $H_1 : \boldsymbol{\theta}_2 \neq \boldsymbol{\theta}_{2,0}$, for a given $\boldsymbol{\theta}_{2,0}$.

For the estimation, we consider both ML and robust methods yielding M-estimators defined as the solution in $\boldsymbol{\theta}$ to

$$\frac{1}{n}\sum_{i=1}^{n} \Psi(\boldsymbol{z}_i; \boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \boldsymbol{0}, \tag{2}$$

where the function Ψ can either be Ψ_{ML} or Ψ_{Robust}, leading respectively to the NB ML estimator (MLE) and its robust counterpart, see Section A in the Supplementary Material. See also Section 3.4 for the regularity conditions we impose on a general Ψ for the relative error properties of the studied tests to be verified. Hereafter, we refer to the M-estimator defining function Ψ as a score, although it need not correspond to a ML score.