Statistics and Probability Letters 106 (2015) 295–300

Contents lists available at ScienceDirect

Statistics and Probability Letters

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

A proportional score test over the nuisance parameter space: Properties and applications

Olivier Thas [a,b], Ao Yuan [c], Hon Keung Tony Ng [d,*], Gang Zheng [1]

[a] Department of Mathematical Modelling, Statistics and Bioinformatics, Ghent University, 9000 Gent, Belgium

[b] National Institute for Applied Statistics Research Australia, School of Mathematics and Applied Statistics, University of Wollongong, New South Wales 2522, Australia

[c] Department of Biostatistics, Bioinformatics and Biomathematics, Georgetown University, Washington, DC 20057, USA

[d] Department of Statistical Science, Southern Methodist University, Dallas, TX 75275, USA

ARTICLE INFO

Article history:
Received 5 February 2015
Received in revised form 20 July 2015
Accepted 20 July 2015
Available online 29 July 2015

Keywords:
BASE
Non-standard hypothesis testing
MAX
Nuisance parameter
Robust test
Score statistic

ABSTRACT

We generalize and study the properties of a test proposed by Zheng (2008) for hypothesis testing involves nuisance parameters that are not present under the null hypothesis. The methodology is illustrated by case-control genetic association studies under model uncertainty.

1. Introduction

In many problems of hypothesis testing, the data-generating model contains nuisance parameters present only under the alternative hypothesis (e.g., Gastwirth, 1966, 1985; Birnbaum and Laska, 1967; Davies, 1977, 1987; Andrews and Ploberger, 1994, 1995; Freidlin et al., 1999; Zhu and Zhang, 2006; Zheng et al., 2009). In these non-standard situations, the nuisance parameters are not identifiable and cannot be estimated consistently under the null hypothesis. As a general model, suppose the data, $X_n = (x_1, \ldots, x_n)$, is a random sample with density function $f(x; \lambda, \boldsymbol{\eta}, \boldsymbol{\theta})$, where $\lambda \in \Lambda \subset R^1$ is the parameter of interest, $\boldsymbol{\eta} \in \boldsymbol{\Omega} \subset R^k$ $(0 \leq k < \infty)$ is a set of nuisance parameters that can be consistently estimated under the null hypothesis, and $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subset R^d$ $(1 \leq d < \infty)$ is a set of nuisance parameters not present under the null hypothesis. We are interested in testing the null hypothesis $H_0 : \lambda = 0$ against the alternative hypothesis $H_1 : \lambda \neq 0$. Under H_0, $f(x; 0, \eta, \theta) = f(x; \eta)$, where θ vanishes and f is known. Note that we restrict the parameter of interest, λ, to be univariate for the sake of simplicity and notational convenience. However, the results developed in this manuscript can be applied to finite-dimensional parameters of interest.

When $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ is given, the score test statistic $Z_n(\boldsymbol{\theta})$ can be derived for testing H_0 (Davies, 1977, 1987). In practice, however, the true value of θ is rarely known and using $Z_n(\boldsymbol{\theta}_0)$ is not robust and can be problematic if $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}$ is misspecified (Gastwirth, 1966, 1985; Freidlin et al., 1999). There are special situations wherein $\boldsymbol{\theta}$ has no impact in the hypothesis testing procedure.

* Corresponding author.
E-mail address: ngh@mail.smu.edu (H.K.T. Ng).

[1] Deceased author.

http://dx.doi.org/10.1016/j.spl.2015.07.022
0167-7152/

For example, when analyzing two-sample ordered categorical data, Kimeldorf et al. (1992) considered a t-test, $t_n(\theta)$, as a function of a one-dimensional score θ assigned to the ordered effect. When $t_n(\theta)$ is always significant (not significant) at the significance level $\alpha = 0.05$, regardless of the value of θ, one always rejects (fails to reject) the null hypothesis of no association at that level regardless of θ. Zheng (2003) applied this idea to testing case-control genetic association with an unknown genetic model indexed by $\theta \in [0, 1]$ and derived algorithms to find when an association test, as a function of θ, is always significant or not for all $\theta \in [0, 1]$.

In practice, given the level α, it is common that the test based on $Z_n(\boldsymbol{\theta})$ is significant for some $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ but not for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. In this case, one can consider the maximin efficiency robust test, which is often a linear combination of $Z_n(\boldsymbol{\theta})$ for several $\boldsymbol{\theta}$ values (Gastwirth, 1966, 1985; Birnbaum and Laska, 1967) or maximum-type tests, for example, MAX $= \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} Z_n^2(\boldsymbol{\theta})$ (Davies, 1987). Alternatively, for the analysis of ordered categorical data, another summary statistic has been considered by Zheng (2008), who studied how often $t_n(\theta)$ is significant at the level α over the parameter space of θ. This summary statistic was referred to as BASE (Zheng, 2008). Although BASE was proposed as a robust test for the ordered categorical data, its properties have not been studied.

In this paper, we give a more general and formal definition of the BASE statistic and study its basic mathematical characteristics and statistical properties for testing hypotheses with non-identifiable nuisance parameters under the null hypothesis. Applications for case-control genetic association studies are presented to illustrate the use of BASE. Comparison of BASE with MAX is reported in the simulations and applications.

2. The score test and BASE

2.1. Notation and the score test

We consider the general setting as described in Section 1. We denote $l(\boldsymbol{x}|\lambda, \boldsymbol{\eta}, \boldsymbol{\theta}) = \log f(\boldsymbol{x}; \lambda, \boldsymbol{\eta}, \boldsymbol{\theta})$, $l_n(\lambda, \boldsymbol{\eta}|\boldsymbol{\theta}) = \sum_{i=1}^n l(x_i|\lambda, \boldsymbol{\eta}, \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is treated as fixed, and $l_n(0, \boldsymbol{\eta}|\boldsymbol{\theta}) = l_n(\boldsymbol{\eta})$. We further denote the first- and second-order partial derivatives of the likelihood function with respect to λ and $\boldsymbol{\eta}$ as $l^{(u,0)}(\boldsymbol{x}|\lambda, \boldsymbol{\eta}, \boldsymbol{\theta}) = \frac{\partial^u}{\partial\lambda^u} l(\boldsymbol{x}|\lambda, \boldsymbol{\eta}, \boldsymbol{\theta})$ for $u = 1, 2$, $l^{(0,1)}(\boldsymbol{x}|\lambda, \boldsymbol{\eta}, \boldsymbol{\theta}) = \frac{\partial}{\partial\boldsymbol{\eta}^T} l(\boldsymbol{x}|\lambda, \boldsymbol{\eta}, \boldsymbol{\theta})$, $l^{(0,2)}(\boldsymbol{x}|\lambda, \boldsymbol{\eta}, \boldsymbol{\theta}) = \frac{\partial^2}{\partial\boldsymbol{\eta}\partial\boldsymbol{\eta}^T} l(\boldsymbol{x}|\lambda, \boldsymbol{\eta}, \boldsymbol{\theta})$ and $l^{(1,1)}(\boldsymbol{x}|\lambda, \boldsymbol{\eta}, \boldsymbol{\theta}) = \frac{\partial^2}{\partial\lambda\partial\boldsymbol{\eta}^T} l(\boldsymbol{x}|\lambda, \boldsymbol{\eta}, \boldsymbol{\theta})$. Then, $l_n^{(u,v)}(\lambda, \boldsymbol{\eta}|\boldsymbol{\theta})$ can be defined in a similar manner. Let $\widehat{\boldsymbol{\eta}}$ be the value of $\boldsymbol{\eta}$ which maximizes $l_n(\boldsymbol{\eta})$. We denote the score function and the Fisher information for λ as $U_n(\lambda, \boldsymbol{\eta}|\boldsymbol{\theta}) = l_n^{(1,0)}(\lambda, \boldsymbol{\eta}|\boldsymbol{\theta})$ and

$$i_n(\lambda, \boldsymbol{\eta}|\boldsymbol{\theta}) = -\left[l_n^{(2,0)}(\lambda, \boldsymbol{\eta}|\boldsymbol{\theta}) - l_n^{(1,1)}(\lambda, \boldsymbol{\eta}|\boldsymbol{\theta}) \left\{l_n^{(0,2)}(\lambda, \boldsymbol{\eta}|\boldsymbol{\theta})\right\}^{-1} \{l_n^{(1,1)}(\lambda, \boldsymbol{\eta}|\boldsymbol{\theta})\}^T\right],$$

respectively. Then, for a given $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, the score test statistic for testing $H_0 : \lambda = 0$ can be written as

$$Z_n(\boldsymbol{\theta}) = \frac{U_n(0, \widehat{\boldsymbol{\eta}}|\boldsymbol{\theta})}{i_n^{1/2}(0, \widehat{\boldsymbol{\eta}}|\boldsymbol{\theta})}. \tag{1}$$

Under H_0, $Z_n(\boldsymbol{\theta}) \xrightarrow{D} N(0, 1)$ for a given $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. We assume that the nuisance parameter space $\boldsymbol{\Theta}$ can be written as $\boldsymbol{\Theta} = \times_{i=1}^d [a_i, b_i]$ and $-\infty < a_i < b_i < \infty$ are known. All the regularity conditions are given in Appendix A.

2.2. Induced Bernoulli random variable

If $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ is known, we reject H_0 when $|Z_n(\boldsymbol{\theta})| > z_{1-\alpha/2}$ at the level α, where $z_{1-\alpha}$ is the upper $100(1-\alpha)$th percentile of the standard normal distribution. Given $Z_n(\boldsymbol{\theta})$ and α, we define an indicator function $\delta_n(\boldsymbol{\theta}, \alpha) = \delta(|Z_n(\boldsymbol{\theta})| > z_{1-\alpha/2})$. Given $\boldsymbol{\theta}$ and α, $\delta_n(\boldsymbol{\theta}, \alpha)$ is a Bernoulli random variable with $\lim_{n\to\infty} \Pr_{H_0}(\delta_n(\boldsymbol{\theta}, \alpha) = 1) = \alpha$ and, as $n \to \infty$, $\mathrm{E}_{H_0}(\delta_n(\boldsymbol{\theta}, \alpha)) \to \alpha$ and $\mathrm{Var}_{H_0}(\delta_n(\boldsymbol{\theta}, \alpha)) \to \alpha(1-\alpha)$. Moreover, given α, $\delta_n(\boldsymbol{\theta}, \alpha)$ is a Bernoulli process indexed by θ, which will be used in Section 3 to approximate the BASE defined in the next subsection.

2.3. Definition of BASE

Let $(\boldsymbol{\Theta}, \mathcal{B}, m)$ be a Lebesgue measurable nuisance parameter space. Assumption A1 of Appendix A implies that, given $\boldsymbol{X}_n = (x_1, \ldots, x_n)$, $Z_n(\boldsymbol{\theta}) : \boldsymbol{\theta} \in \boldsymbol{\Theta} \mapsto \mathcal{R}$ is a smooth function with continuous first derivative for any $\boldsymbol{\theta} \in \times_{i=1}^d (a_i, b_i)$. For the weak convergence in Assumption A2 of Appendix A, we need to specify a metric on the function space $\mathcal{F}$ of $\mathcal{R}^h$-valued functions with $h < \infty$ on $\boldsymbol{\Theta}$. We assume that the metric is chosen so that the function $U(\cdot) \to m\{\boldsymbol{\theta} \in \boldsymbol{\Theta} : U(\boldsymbol{\theta}) > c\}$ is continuous at each function $U \in \mathcal{F}$ that is continuous on $\boldsymbol{\Theta}$, where m is a Lebesgue measure and c is a constant. This condition holds, for example, if the uniform or Skorohod metric is used. The excursion set $\left\{\boldsymbol{\theta} \in \boldsymbol{\Theta} : |Z_n(\boldsymbol{\theta})| \geq z_{1-\alpha/2}\right\}$ is denoted by $S_n(\boldsymbol{\Theta}, z_{1-\alpha/2})$, which is $\mathcal{B}$-measurable given X_n. The BASE statistic is then defined as $m\left\{S_n(\boldsymbol{\Theta}, z_{1-\alpha/2})\right\}$, the Lebesgue measure of $S_n(\boldsymbol{\Theta}, z_{1-\alpha/2})$. Note that the definition given here generalizes the BASE defined in Zheng (2008).