Statistics and Probability Letters 116 (2016) 94–100

Contents lists available at ScienceDirect

Statistics and Probability Letters

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

Asymmetric risk of the Stein variance estimator under a misspecified linear regression model

Huaizhen Qin *, Weiwei Ouyang

[a] Center for Bioinformatics and Genomics, Tulane University School of Public Health and Tropical Medicine, 1440 Canal Street, New Orleans, LA 70112, United States

[b] Department of Biostatistics and Bioinformatics, Tulane University School of Public Health and Tropical Medicine, 1440 Canal Street, New Orleans, LA 70112, United States

ARTICLE INFO

Article history:
Received 24 October 2015
Received in revised form 20 April 2016
Accepted 21 April 2016
Available online 28 April 2016

MSC:
primary 62J99
secondary 62G15

Keywords:
LINEX loss
Dominance
Inadmissibility
Pre-test variance estimator
Latent biomarkers

ABSTRACT

Using LINear EXponential (LINEX) loss, we formulate the risk of a generic pre-test estimator for the disturbance variance in a misspecified regression model. The dominance and inadmissibility of the Stein estimator of disturbance variance are derived under respective loss configures.

1. Introduction

In real-world regression analyses, important predictors are often unintentionally omitted. For instance, complex human diseases are influenced by many genetic variants, mediate biomarkers, environmental covariates, and their interactions (McCarthy et al., 2008; Thomas, 2010). Existent regression analyses in modern multi-omics studies often unavoidably omit important causal variants, biomarkers and covariates, either observed or latent or both. Model misspecifications occur frequently and widely, as driven by complexity of systems biology, simplicity of working models adopted for data analyses, and more. To localize causal loci of a quantitative trait, a typical single-marker association test is based on the least squares regression of the trait on several observed covariates and genotypic score of one single genomic marker at a time (Churchill and Doerge, 1994). Such a standard test omits all the other causal loci even though their genotypes are available in the dataset. To date, however, little has been done to formulate the consequences of model misspecifications in massive genetic association tests. One key task is to formulate the impacts of model misspecifications on the risks of disturbance variance estimators.

In the context of a Gaussian disturbance, the Stein variance estimator dominates the usual variance estimator in terms of mean squared error (Stein, 1964). The Stein variance estimator in a linear regression model can be viewed as a pre-test

* Correspondence to: Department of Biostatistics and Bioinformatics, Tulane University School of Public Health and Tropical Medicine, 1440 Canal Street, Suite 2001 New Orleans, LA 70112, United States. Tel.: +1 504 988 2042; fax: +1 504 988 1706.

E-mail addresses: hqin2@tulane.edu (H. Qin), wouyang@tulane.edu (W. Ouyang).

http://dx.doi.org/10.1016/j.spl.2016.04.019

estimator (Gelfand and Dey, 1988). Mean squared errors of disturbance variance estimators have been examined under multiple regression models when there are restrictions on regression coefficients (Clarke et al., 1987a,b) and when there are omitted variables (Giles and Clarke, 1989). In terms of mean squared error, the pre-test variance estimator dominates the usual estimator if the critical value of the pre-test is chosen appropriately, even some variables are omitted (Giles, 1991). The dominance of the Stein estimator over the usual estimator has proven robust to the degree of model misspecification (Ohtani, 2002).

All aforesaid investigations are based on quadratic loss, which is a typical symmetric loss that imposes equal penalty on over- and under-estimations. There are situations where over- and under-estimations can lead to different consequences. Some typical examples illustrated that the quadratic loss can be unduly restrictive and inappropriate (Wan et al., 2000). Thus, it can be instructive to investigate the risk properties of estimators under asymmetric losses. Since Varian (1975), several authors have considered LINEX loss in various problems of interest. Some examples are Zellner (1986), Cain and Janssen (1995), Zou (1997) and Wan and Kurumai (1999), among others. It is unclear if the Stein variance estimator still robustly dominates the usual estimator in terms of asymmetric losses, e.g., the LINEX loss.

The rest of the article is arranged as the following. In Section 2, we formulate the LINEX loss, model misspecification, and generic pre-test estimator family of disturbance variance. In Section 3, we mathematically investigate the risk performance of the generic pre-test estimator under a positively scaled LINEX loss and establish the dominance of the Stein estimator over a large class of pre-test variance estimators. In Section 4, we numerically investigate the risk performance of the generic pre-test estimator under a negatively scaled LINEX loss to demonstrate the inadmissibility of Stein variance estimator. In Section 5, we conclude this article according to our theoretical and numerical findings. We present mathematical details in the Appendix.

2. Model formulation

Let $\hat{\sigma}^2$ be an arbitrary estimator for disturbance variance σ^2. The generic LINEX loss of $\hat{\sigma}^2$ can be defined as

$$L(\hat{\sigma}^2, \sigma^2) = r\{\exp(a\Delta) - a\Delta - 1\}, \tag{1}$$

where $\Delta = \hat{\sigma}^2/\sigma^2 - 1$, $a \neq 0$ is a shape parameter, and $r > 0$ is a factor of proportionality. The sign of a reflects the direction of asymmetry. If $a > 0$ ($a < 0$), then over-(under-) estimation is considered to be more serious than under-(over-) estimation of the same degree. The magnitude of a reflects the degree of asymmetry. Loss (1) approximately collapses to the quadratic loss $\frac{1}{2}ra^2\Delta^2$ for small values of a. In such a sense, the LINEX loss can be regarded as a generalization of the squared loss allowing for asymmetry.

Suppose the observed data $[y, X]$ and latent independent variables follow linear regression model

$$y = X\beta + Z\gamma + \varepsilon, \tag{2}$$

where y is an $n \times 1$ vector of observations on a dependent variable; X is an $n \times k$ matrix of observed data of non-stochastic independent variables; and Z is an $n \times l$ matrix of l latent non-stochastic independent variables; β and γ are $k \times 1$ and $l \times 1$ vectors of regression coefficients; ε is an $n \times 1$ vector of random errors with $\mathrm{E}[\varepsilon] = 0$, $\mathrm{E}[\varepsilon\varepsilon'] = \sigma^2 I_n$, and I_n stands for the identity matrix of order n. We assume that X and Z are of full column ranks. Although the correct model is (2), in practice the matrix Z may be omitted mistakenly, and the model is misspecified as

$$y = X\beta + \eta, \tag{3}$$

where $\eta = Z\gamma + \varepsilon$ is mistakenly taken as the vector of random errors. The OLS estimator for β based on the misspecified model is $b = S^{-1}X'y$, where $S = X'X$. The residual vector is $e = y - Xb$ and the usual estimator for σ^2 is

$$\hat{\sigma}^2 = \frac{e'e}{v+2}, \tag{4}$$

where $v = n - k$. In terms of MSE, it is well known that the usual estimator $\hat{\sigma}^2$ dominates the estimator $e'e/v$, and the Stein estimator

$$\hat{\sigma}_S^2 = \min\left\{\frac{y'y}{n+2}, \frac{e'e}{v+2}\right\}, \tag{5}$$

dominates the usual estimator $\hat{\sigma}^2$ (Ohtani, 2002).

In this article, we inspect the dominance and inadmissibility of the Stein estimator by formulating the LINEX-risk of the following generic pre-test estimator for σ^2:

$$\hat{\sigma}^2(\tau) = I(F \le \tau)\frac{y'y}{n+2} + I(F > \tau)\frac{e'e}{v+2}, \tag{6}$$

where $I(E)$ is the indicator function such that $I(E) = 1$ if the event E occurs and $I(E) = 0$ otherwise, $F = (b'Sb/k)/(e'e/v)$ is the test statistic for the null hypothesis $\mathrm{H}_0 : \beta = 0$ versus the alternative hypothesis $\mathrm{H}_1 : \beta \neq 0$, and τ is the critical