Computational Statistics and Data Analysis 53 (2009) 2537–2549

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Shrinkage estimation in general linear models

Lihua An [a,b,c,*], Sévérien Nkurunziza [b], Karen Y. Fung [b,a], Daniel Krewski [a], Isaac Luginaah [d]

[a] McLaughlin Centre for Population Health Risk Assessment, University of Ottawa, Canada
[b] Department of Mathematics and Statistics, University of Windsor, Canada
[c] Household Survey Methods Division, Statistics Canada, Canada
[d] Department of Geography, University of Western Ontario, Canada

ARTICLE INFO

Article history:
Received 24 March 2008
Received in revised form 30 November 2008
Accepted 30 November 2008
Available online 7 December 2008

ABSTRACT

We propose a James–Stein-type shrinkage estimator for the parameter vector in a general linear model when it is suspected that some of the parameters may be restricted to a subspace. The James–Stein estimator is shown to demonstrate asymptotically superior risk performance relative to the conventional least squares estimator under quadratic loss. An extensive simulation study based on a multiple linear regression model and a logistic regression model further demonstrates the improved performance of this James–Stein estimator in finite samples. The application of this new estimator is illustrated using Ontario newborn infants data spanning four fiscal years.

1. Introduction

In many applications, there may be uncertain prior information available about the parameters in the statistical model used to describe the available data. Inferences about these parameters may be improved if the uncertain prior information is incorporated properly in the estimation procedure. In clinical trials involving new drugs, for example, results from several similar trials may be pooled to provide a sufficiently large enough data to support reliable estimates and statistical conclusions. In statistical analyses of spontaneous reports of adverse drug reactions, logistic regression models are often used to predict the likelihood of occurrence of a particular adverse event when a drug, or combination of drugs, is taken. If it assumed that the drugs affect individuals in different countries in the same way, a pooled analysis of data from those countries may be undertaken, leading to a pooled estimate which will be more precise than the country-specific estimates. However, if this prior assumption of similar effects of the drug(s) of interest is not true, the pooled estimator may be subject to substantial bias, and greater uncertainty than is suggested by the pooled analysis. As a second example, in studying risk factors associated with low infant birth weight, a linear regression model using factors such as maternal age, height, smoking status, or presence of hypertension as independent variables can be fit for different racial group. If we suspect that the regression equations are similar for different racial groups, a common regression model can be fitted for all racial groups. Since the validity of the prior assumption is not tested, neither the pooled nor unrestricted estimators make use of the available information in an optimal way. The James–Stein-type shrinkage estimator incorporates this uncertain prior information, and combines the restricted and unrestricted estimators in a superior manner. In this paper, the performance of James–Stein-type shrinkage estimator is explored, and its use is illustrated using both actual and simulated data sets.

* Corresponding address: Household Survey Methods Division, Statistics Canada, R.H.Coats Building, 16-A 100 Tunney's Pasture Driveway, Ottawa, Ontario, Canada K1A 0T6. Tel.: +1 613 951 7248; fax: +1 613 951 3100.
E-mail address: lihua.an@statcan.gc.ca (L. An).

0167-9473/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2008.11.027

Formally, we consider the general linear model problem of estimating the $p = (k+1)$-dimensional parameter vector $\boldsymbol{\beta}$ when one observes an n-dimensional sample vector $\mathbf{y}$, such that

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}. \tag{1.1}$$

Here, $\mathbf{X}$ is an $(n \times p)$ design matrix of k independent variables and $\boldsymbol{\varepsilon}$ is the error term with $E(\boldsymbol{\varepsilon}) = \mathbf{0}$ and $\text{cov}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}_n$. The scale parameter σ^2 is usually unknown. The objective is to estimate the unknown vector $\boldsymbol{\beta}$ by an estimator $\boldsymbol{\delta}(\mathbf{y})$ when the performance is evaluated by a quadratic risk measure

$$\mathfrak{R}(\boldsymbol{\delta}(\mathbf{y}), \boldsymbol{\beta}) = E\{(\boldsymbol{\delta}(\mathbf{y}) - \boldsymbol{\beta})'\mathbf{W}(\boldsymbol{\delta}(\mathbf{y}) - \boldsymbol{\beta})\},$$

where $\mathbf{W}$ is any positive semi-definite matrix. Assuming the usual regularity conditions underlying the general linear model, the conventional least squares (LS) estimator is $\boldsymbol{\delta}^{LS}(\mathbf{y}) = \hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$. Under quadratic loss, $\boldsymbol{\delta}^{LS}(\mathbf{y})$ is a minimax estimator with constant risk $\mathfrak{R}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) = \sigma^2\text{trace}(\mathbf{W}(\mathbf{X}'\mathbf{X})^{-1})$. The properties of this estimator have been discussed in text books, such as Lehmann and Casella (1998).

If the parameters given in (1.1) are restricted to a subspace

$$C\boldsymbol{\beta} = d, \tag{1.2}$$

where C is a matrix and d is a vector, the LS estimator $\boldsymbol{\delta}^{LS}(\mathbf{y})$ can be replaced by the restricted estimator $\hat{\boldsymbol{\beta}}^{re} = \boldsymbol{\delta}^{LS}(\mathbf{y}) - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}'[\mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}']^{-1}\left(C\boldsymbol{\delta}^{LS}(\mathbf{y}) - d\right)$. However, in practice, the suspected relation (1.2) needs to be tested and confirmed based on actual data. Here, we adopt the James–Stein-type shrinkage method in order to develop a test statistic that incorporates prior uncertainty about the constraint in (1.2), and then combine the unrestricted and restricted estimates in an optimal way to achieve an improved estimator of the model parameters.

The history of shrinkage estimation dates back to the initial work by Charles Stein in 1956 (Stein, 1956). Since then considerable research has been done focusing on shrinkage estimation of location parameters (c.f., Strawderman (1971), Sen and Saleh (1985), Casella and Hwang (1986), Saleh and Sen (1987) and Ahmed and Saleh (1999)). Bayesian approaches to obtain similar estimators have been explored by Efron and Morris (1973), Berger (1982), and Kempthorne (1988), among others. Green and Strawderman (1991) showed how to combine biased and unbiased estimators using the Stein shrinkage concept. Kim and White (2001) extended the James–Stein-type estimators to the Least Absolute Deviations Estimator. Spitzner (2005) extended the benefits of risk-reducing shrinkage estimation from normal theory model to arbitrary GLMs. Saleh (2006) provided a comprehensive discussion of Stein-type estimators. Additionally recent work has been reported by Ahmed and Krzanowski (2004), Ahmed (2005), Khan and Ahmed (2006), Zoua et al. (2007), and Pardoa and Menéndez (2008), among others. In this paper, we extend the shrinkage estimation method to the general linear model under any general linear hypothesis about the underlying model parameters and examine its asymptotic properties as well as finite sample performance using computer simulation.

The remainder of this paper is organized as follows. In Section 2, we present the James–Stein-type estimator. A comparison of unrestricted and shrinkage estimators is provided in Section 3. In Section 4, we examine the finite sample properties of the estimators using computer simulation. Two illustrative examples are provided in Section 5. Our conclusions are presented in Section 6. Proofs of the propositions stated in Section 3 are provided in the Appendix.

2. James–Stein-type estimator

Suppose we have q independent regression equations based on q independent data sets of sizes $n_1, n_2, \ldots, n_q$. The q independent models can be combined into a single model of the general form given in Eq. (1.1), where $\mathbf{y}$ is a vector of length n, $n = \sum_{i=1}^{q} n_i$, $\mathbf{X}$ is an $n \times pq$ design matrix, and $\boldsymbol{\beta}$ is a vector of length pq. Suppose it is suspected that some of the parameters may be restricted to a subspace. One can test the null hypothesis $H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{d}$, where $\mathbf{C}$ is an $r \times pq$ matrix of rank r and $r \leq pq$. This general linear null hypothesis allows statistical tests to be conducted on any subset of the model parameters. One test of particular interest is whether the slopes of the q regression lines are equal. When $k = 1$, the null hypothesis can be written as

$$H_0 : \begin{bmatrix} 0 & 0 & \cdots & 0 & 1 & -1 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 1 & 0 & -1 & \cdots & 0 \\ \vdots & & & \vdots & & \vdots & & & \\ 0 & 0 & \cdots & 0 & 1 & 0 & 0 & \cdots & -1 \end{bmatrix} \begin{bmatrix} \beta_{01} \\ \beta_{02} \\ \vdots \\ \beta_{0q} \\ \beta_{11} \\ \beta_{12} \\ \vdots \\ \beta_{1q} \end{bmatrix} = \mathbf{0}.$$

The test statistic is given by $F = \frac{[\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{d}]'[\mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}']^{-1}[\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{d}]}{rs^2}$, where s^2 is the residual mean squares under the full model.

Let us denote the least square estimates under the unrestricted model and restricted model (under H_0) by $\hat{\boldsymbol{\beta}}^{un}$ and $\tilde{\boldsymbol{\beta}}^{re}$, respectively. If H_0 is true, the restricted estimator $\tilde{\boldsymbol{\beta}}^{re}$ is expected to be a better estimator since it pools samples together and borrows strength across q data sets. However, when the null hypothesis is not true, the restricted or pooled estimator may