Biostatistics and Biometrics
Open Access Journal
ISSN: 2573-2633

Juniper Publishers
key to the Researchers

Review Article
Volume 9 Issue 2 - February 2019
DOI: 10.19080/BBOAJ.2019.09.555756

Biostat Biometrics Open Acc J

A Note on Preliminary Test Estimator in High Dimensional Regression Model

Arashi M* and Norouzirad M

Department of Statistics, Shahrood University of Technology, Iran

Submission: October 06, 2018; **Published:** February 25, 2019

***Corresponding author:** M. Arashi, Department of Statistics, Faculty of Mathematical Science, Shahrood University of Technology, Shahrood, Iran

Abstract

In this note, we show how to construct the well-known preliminary test estimator to be practical in the high dimensional linear regression model. The strategy is to use a high dimensional test when it is a priori suspected that the parameter may be equal to zero.

Keywords: High dimension; Preliminary test estimator; Regression model

Abbreviations: UE: Unrestricted Estimator; PTE: Preliminary Test Estimator; DLBCL: Diffuse Large B-Cell Lymphoma; CLL: Chronic Lymphocytic Leukemia; FL: Follicular Lymphoma

Introduction

In genomic studies the collected data have a challenging feature that the number of variables is relatively large to the number of sample. Genetic epidemiology, apparently is a suitable example of such data sets, where the number of subjects, n, is in the thousands, while the number of variables, p, ranges from tends of thousands to hundreds of thousands of genetic features. In the high dimensional regression model three goals may be considered. According to Wasserman and Roeder [1], the first and second are finding models with good prediction error and estimating the true sparsity pattern. The second goal deals with the set of covariates $D=\{j:\beta_j\neq 0\}$ with non-zero regression coefficient. However, there is another goal that has attracted considerable attention among researchers. It is refining the sample covariance matrix in some sense, that can be invertible to construct the test statistic and do the related inferential issues. This goal does not incorporate variable selection methods to map the data into a subset of $S=\{1,2,\ldots,p\}$.

The methodology that we use to follow the third goal is adopted from Srivastava [2]. The idea is that in high dimensional setup, since n is smaller than p, the sample covariance matrix, S, may be singular, however the diagonal matrix constructing from diagonal elements of S is non-singular and the natural inferential analysis can be considered using this diagonal matrix under some mild regularity conditions.

In this paper, we deem to extend the estimation theory of mean parameter in multivariate regression models by incorporating non-sample prior information under a high dimensional setup.

High Dimensional Preliminary Test Estimator

Consider the high dimensional regression model given by

$$Y_\alpha=\theta+\varepsilon_\alpha;\qquad \alpha=1,2,\ldots,N,\ (2.1)$$

where, $Y_\alpha=(Y_{\alpha 1},\ldots,Y_{\alpha p})^T$ is the p-vector response, with $p>N$, $\varepsilon_\alpha=(\varepsilon_{\alpha 1},\ldots,\varepsilon_{\alpha p})^T$ is the p-vector of disturbances, $\theta=(\theta_1,\ldots,\theta_p)^T$ is the location vector. We do not assume any specific distributional assumption for the error term, however it has the following characteristics,

$$E(\varepsilon_\alpha)=0,\quad E(\varepsilon_\alpha\varepsilon_\alpha^T)=\Sigma,\quad \alpha=1,2,\ldots,N.$$

The unrestricted estimator (UE) of θ is given by

$$\tilde{\theta}_N=\bar{Y}=\frac{1}{N}1_N^T\begin{pmatrix}Y_1\\ \vdots\\ Y_N\end{pmatrix}=\begin{pmatrix}\bar{Y}_1\\ \vdots\\ \bar{Y}_N\end{pmatrix}.$$

Define the matrix $S=\frac{1}{n}V,\ n=N-1,$ where

$$V=\sum_{\alpha=1}^{N}(Y_\alpha-\bar{Y})(Y_\alpha-\bar{Y})^T.$$

In the high dimensional regression, the matrix V may be singular. However, the diagonal matrix consisting of only the diagonal elements of $V=(v_{ij})$, defined as

$$D=\mathrm{Diag}(s_{11},\ldots,s_{pp})$$

is a non-singular matrix. To construct the test statistics for testing $H_0:\theta=0$ vs. $H_A:\theta\neq 0,$ we use the test statistic

$$T=\frac{N\tilde{\theta}_N^T D^1\tilde{\theta}_N-\frac{np}{n-2}}{\left(2\mathrm{tr}\left(R^2\right)-\frac{p^2}{n}\right)^{\frac{1}{2}}},\quad R=D^{-\frac{1}{2}}S\,D^{-\frac{1}{2}}$$

We reject the null-hypothesis as soon as $T\geq Z_{1-\alpha}$, where $Z_{1-\alpha}$ is the upper α critical value from the distribution of T.

We make use of the following result for the distribution of the test statistic T.

Theorem 1. [2]

Assume for some $\alpha\in(0,1]$, $n=O(p^{\alpha})$ and for $i=1,2,3,4$, $\lim_{p\to\infty}\left(\frac{\mathrm{tr}(R^i)}{p}\right)=\tau_{i0}<\infty$, where $R=D^{-\frac{1}{2}}\Sigma D^{-\frac{1}{2}}$, $\Sigma=(\sigma_{ij})$, $D^{-\frac{1}{2}}=Diag\left(\sigma_{11}^{-\frac{1}{2}},\ldots,\sigma_{pp}^{-\frac{1}{2}}\right)$. Then under the null-hypothesis $H_0:\theta=0$, the test statistics T has the following asymptotic distribution

$$\lim_{(n,p)\to\infty}P(T<z)=\Phi(Z),$$

Where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution. Further, suppose that under the alternative hypothesis H_A, the local assumption $\theta=\left(\frac{1}{nN}\right)^{\frac{1}{2}}\delta$, for some constant vector δ holds, Then the test statistics T has the following asymptotic distribution.

$$\lim_{(n,p)\to\infty}P(T<z)=\Phi\left(Z+\frac{\delta^T D^{-1}\delta}{n\sqrt{2tr\left(R^2\right)}}\right),$$

where, $\frac{\delta^T D^{-1}\delta}{p}\leq M$ for some M, which does not depend on p.

Following Saleh [3], the high dimensional preliminary test estimator (PTE), is defined as

$$\hat{\theta}_N^{PT}=\tilde{\theta}_N I(T\geq z_{1-\alpha})+\hat{\theta}_N I(T<Z_{1-\alpha})=\tilde{\theta}_N-\tilde{\theta}_N I(T<Z_{1-\alpha})\quad(2.2)$$

where $\hat{\theta}_N=0$ is the restricted estimator (RE).

Vision of the risk

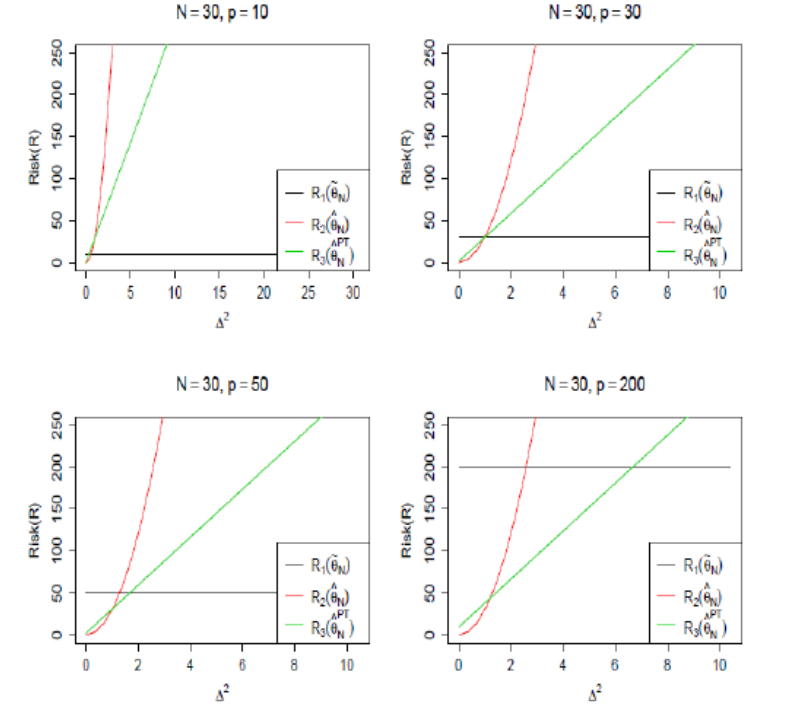

Figure 1: An asymptotic view about the risk of the estimators in the simulation.

In this section, we provide a graphical illustration of the risk function, from the asymptotic view point. We display the graphs of the risk functions for the three proposed estimator. For our purpose, we assume $\Sigma=I_p$, $N=30$ and $p\in(10,30,50,200)$, $n=1-\alpha$ and $\delta=1_p$. Consider that for the case $D=I_p$, the regularity conditions of Theorem 1 satisfy. Since the error term in the multivariate model is distribution free, we prefer not to generate Y and therefore, to incorporate a singular matrix V in our computation we generate $(p-1)\times p$ random matrix and let the p^{th} row be the sum of the others. From Figure 1 it is clear that the risk functions of the RE and PTE have increasing trend with respect to $\Delta^2=\theta^T\theta$. It can be also seen as p increases, domination nature of the PTE over the UE covers for a bigger range of Δ^2, which again suggests the superiority of this estimator over the UE.

Blood Cancer Data

The lymphoma leukemia dataset provides expression for $p=4,026$ genes across $N=62$ patients with 42 samples of diffuse large B-cell lymphoma (DLBCL), 9 samples of follicular lymphoma (FL), and 11 samples of chronic lymphocytic leukemia (CLL). This data is taken from the package "spls" [4]. Lymphoma is a type of cancer that begins in immune system cells called lymphocytes. Like

How to cite this article: Arashi M, Norouzirad M. A Note on Preliminary Test Estimator in High Dimensional Regression Model. Biostat Biometrics Open Acc J. 2019; 9(2): 555756. DOI: 10.19080/BBOAJ.2019.09.555756.

other cancers, lymphoma occurs when lymphocytes are in a state of uncontrolled cell growth and multiplication. Lymphocytes are white blood cells that move throughout the body in a fluid called lymph. Also, lymphocytes Lymphoma is the most common form of hematological malignancy, or "blood cancer", in the developed world. The value of the test statistics is equal to 4.972658 that compared to $z_{0.95}=1.96,$ suggests that the PTE is reduced to the UE, from an asymptotic viewpoint.

References

1. Wasserman L, Roader K (2009) High-dimensional variable selection. Ann Statist 37(5A): 2178-2261.
2. Sirvastava MS (2009) A test for the mean vector with fewer observation than the dimension under non-normality. J Mult Anal 100(3): 518-532.
3. Saleh AK Md Ehsanes (2006) Theory of Preliminary Test and Stein-Type Estimation with Applications, John Wiley, New York.
4. Chung D, Chun H, Keles S, (2012) spls: Sparse Partial Least Squares (SPLS) Regression and Classification. (R package version 2.1-2).

DOI: 10.19080/BBOAJ.2019.09.555756

How to cite this article: Arashi M, Norouzirad M. A Note on Preliminary Test Estimator in High Dimensional Regression Model. Biostat Biometrics Open Acc J. 2019; 9(2): 555756. DOI: 10.19080/BBOAJ.2019.09.555756.