Extending a Large Dimensional Empirical Likelihood Test to Regression

Ye Alex Zhao*, Changcheng Li and Runze Li

yazhao@psu.edu

Department of Statistics, Pennsylvania State University

Abstract

Using empirical likelihood methods for inference on a population mean when working with a high dimensional vector (where the dimension p is greater than the sample size n) has generally faced difficulties both because the convex hull of the observations becomes too small to cover the true mean value and because the sample covariance matrix is singular. Recent research has proposed a new strategy that addresses these two issues. An extension of this strategy towards hypothesis testing in the linear regression setting is presented, with simulated data used to illustrate the test's power performance.

Introduction

A common problem faced in statistics involves data $\{\boldsymbol{x}_i, y_i\}, i = 1, \ldots, n$ to which is fit a linear regression model

$$y = \boldsymbol{x}^T\boldsymbol{\beta} + \epsilon \tag{1}$$

In these cases, there is interest in testing the hypothesis:

$$H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0 \quad \text{versus} \quad H_1 : \boldsymbol{\beta} \neq \boldsymbol{\beta}_0 \tag{2}$$

- In the cases where the dimension of the data p is smaller than the sample size n, this problem is well defined and has well known solutions (least squares regression, etc)
- However, when $p >> n$, this problem is much harder because
  1. The sample covariance matrix is singular (most commonly used test statistics require an inverse of the covariance matrix in their calculation)
  2. The convex hull of the observations are too small the cover the true mean (need to add artificial datapoints)
- Recent developments [1] have developed test statistics for high-dimensional empirical likelihood settings, which we extend to high-dimensional regression settings

Methods

- In linear regression, the score equation for $\boldsymbol{\beta}$ is

$$\sum_{i=1}^{n} \boldsymbol{x}_i(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}) = \mathbf{0} \tag{3}$$

- For $\boldsymbol{z}_i = \boldsymbol{x}_i(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})$ and $\boldsymbol{\mu} = E(\boldsymbol{z_i})$, testing

$$H_0 : \boldsymbol{\mu} = \mathbf{0} \quad \text{versus} \quad H_1 : \boldsymbol{\mu} \neq \mathbf{0} \tag{4}$$

is equivalent to testing (2).

1. Convex Hull

   Add two synthetic data points:

$$\boldsymbol{z}_{n+1} = -a_n(\bar{\boldsymbol{z}}) \tag{5}$$

$$\boldsymbol{z}_{n+2} = (2 + a_n)(\bar{\boldsymbol{z}}) \tag{6}$$

$$a_n = \frac{l_n}{\{||\bar{\boldsymbol{z}}||^2 + k_n|\alpha^T\bar{\boldsymbol{z}}|^2\}^{1/2}} \tag{7}$$

   With k_n, l_n positive constants and $||\alpha|| = 1$.

$$W(k_n) = -2\ln R(k_n) \tag{8}$$

   With $R(k_n)$ being the empirical likelihood ratio.

2. Covariance Matrix Singularity

   Our test statistic is:

$$T_n = \{2\widehat{\mathrm{tr}(\Omega^2)}\}^{-1/2}\{\frac{2nl_n^2}{(n+2)^2}W(k_n) - \widehat{\mathrm{tr}(\Omega)}\} \tag{9}$$

   With defined terms:

$$\widehat{\mathrm{tr}(\Omega)} = \widehat{\mathrm{tr}(\Sigma)} + k_n\alpha^T\hat{\Sigma}\alpha$$

$$\widehat{\mathrm{tr}(\Omega^2)} = \widehat{\mathrm{tr}((\Sigma^2)} + 2k_n\alpha^T\hat{\Sigma}^2\alpha + k_n^2(\alpha^T\hat{\Sigma}\alpha)^2$$

$$\widehat{\mathrm{tr}(\Sigma)} = (\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{z}_i^T\boldsymbol{z}_i - \frac{1}{P_n^2}\sum_{i\neq j}\boldsymbol{z}_i^T\boldsymbol{z}_i)$$

$$\widehat{\mathrm{tr}(\Sigma^2)} = (\frac{1}{P_n^2}\sum_{i\neq j}(\boldsymbol{z}_i^T\boldsymbol{z}_j)^2 - \frac{2}{P_n^3}\sum_{i,j,k}^{*}\boldsymbol{z}_i^T\boldsymbol{z}_j\boldsymbol{z}_j^T\boldsymbol{z}_k + \frac{1}{P_n^4}\sum_{i,j,k,l}^{*}\boldsymbol{z}_i^T\boldsymbol{z}_j\boldsymbol{z}_k^T\boldsymbol{z}_l)$$

   Where $P_n^r = \frac{n!}{(n-r)!}$, and $\sum^*$ are the summation over mutually different indices.

Under certain assumptions, $T_n \to N(0,1)$ as $n \to \infty$ for H_0.

Results

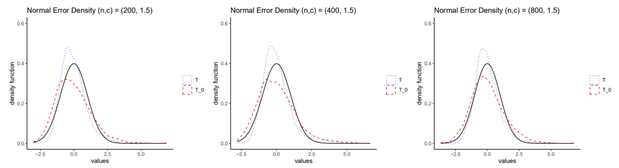

Figure 1:Estimated density curves of T_n and T_n^0 under H_0. The solid, dotted, and dashed curves are the density curves of N(0,1), T_n and T_n^0, respectively.

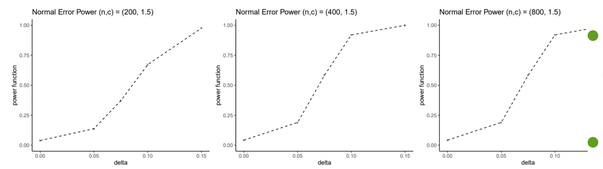

Figure 2:Empirical power functions of T_n with $p/n = 1.5$.

Sensitivity results for changes in α, l_n, and k_n for the linear regression setting are shown in the tables below:

δ	c1	c2	c3
0.000	0.444	0.036	0.042
0.050	0.927	0.271	0.190
0.075	0.999	0.757	0.584
0.100	1.000	0.969	0.920
0.150	1.000	1.000	1.000

Table 1:Empirical power of T_n with c1, c2, and c3 for $\alpha = e_i$ when the ith element of $|\bar{z} - \mu|$ is the largest, α is a random direction with $||\alpha|| = 1$, and $\alpha = 1_p/\sqrt{p}$, respectively

δ	c1	c2	c3
0.000	0.444	0.036	0.042
0.050	0.927	0.271	0.190
0.075	0.999	0.757	0.584
0.100	1.000	0.969	0.920
0.150	1.000	1.000	1.000

Table 2:The empirical power of T_n with d1 and d2 for $l_n = 0.5n^{5/4}\ln(n)$ and $l_n = 2n^{5/4}\ln(n)$, respectively

Linear

δ	c1	c2	c3
0.000	0.444	0.036	0.042
0.050	0.927	0.271	0.190
0.075	0.999	0.757	0.584
0.100	1.000	0.969	0.920
0.150	1.000	1.000	1.000

Table 3:The empirical power of T_n with e1 and e2 for $k_n = 0.5(p/\ln p)^{1/2}$ and $k_n = 2(p/\ln p)^{1/2}$, respectively

Conclusion

- Under the high-dimensional linear regression setting, there exists a method for hypothesis testing on $\boldsymbol{\beta}$ using the empirical likelihood
- This approach addresses issues around the convex hull of the $p >> n$ setting and the singularity of the covariance matrix
- The constructed test statistic has asymptotic normality and good empirical power
- An extension of this work towards logistic and Poisson regression has simulation results (but are omitted here)

References

[1] Xia Cui, Runze Li, Guangren Yang, and Wang Zhou. Empirical likelihood test for a large-dimensional mean vector. *Biometrika*, 2020.

[2] Art B Owen. *Empirical likelihood*. CRC press, 2001.

[3] Jing Chen, Xiaoyi Wang, Shurong Zheng, Baisen Liu, and Ning-Zhong Shi. Tests for high-dimensional covariance matrices. *Random Matrices: Theory and Applications*, page 2050009, 2019.

[4] Zhi-Dong Bai and Yong-Qua Yin. Limit of the smallest eigenvalue of a large dimensional sample covariance matrix. In *Advances In Statistics*, pages 108–127. World Scientific, 2008.