ARTICLE IN PRESS

Journal of Econometrics ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Journal of Econometrics

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

Testing a single regression coefficient in high dimensional linear models

Wei Lan [a,*], Ping-Shou Zhong [b], Runze Li [c], Hansheng Wang [d], Chih-Ling Tsai [e]

[a] Statistics School and Center of Statistical Research, Southwestern University of Finance and Economics, Chengdu, PR China
[b] Department of Statistics and Probability, Michigan State University, East Lansing, MI 48823, United States
[c] Department of Statistics and the Methodology Center, The Pennsylvania State University, University Park, PA 16802-2111, United States
[d] Department of Business Statistics and Econometrics, Guanghua School of Management, Peking University, Beijing, 100871, PR China
[e] Graduate School of Management, University of California, Davis, CA 95616-8609, United States

ARTICLE INFO

Article history:
Received 15 February 2016
Received in revised form
15 February 2016
Accepted 21 May 2016
Available online xxxx

Keywords:
Correlated Predictors Screening
False discovery rate
High dimensional data
Single coefficient test

ABSTRACT

In linear regression models with high dimensional data, the classical z-test (or t-test) for testing the significance of each single regression coefficient is no longer applicable. This is mainly because the number of covariates exceeds the sample size. In this paper, we propose a simple and novel alternative by introducing the Correlated Predictors Screening (CPS) method to control for predictors that are highly correlated with the target covariate. Accordingly, the classical ordinary least squares approach can be employed to estimate the regression coefficient associated with the target covariate. In addition, we demonstrate that the resulting estimator is consistent and asymptotically normal even if the random errors are heteroscedastic. This enables us to apply the z-test to assess the significance of each covariate. Based on the p-value obtained from testing the significance of each covariate, we further conduct multiple hypothesis testing by controlling the false discovery rate at the nominal level. Then, we show that the multiple hypothesis testing achieves consistent model selection. Simulation studies and empirical examples are presented to illustrate the finite sample performance and the usefulness of the proposed method, respectively.

1. Introduction

In linear regression models, it is a common practice to employ the z-test (or t-test) to assess whether an individual predictor (or covariate) is significant when the number of covariates (p) is smaller than the sample size (n). This test has been widely applied across various fields (e.g., economics, finance and marketing) and is available in most statistical software. One usually applies the ordinary least squares (OLS) approach to estimate regression coefficients and standard errors for constructing a z-test (or t-test); see, for example, Draper and Smith (1998) and Wooldridge (2002). However, in a high dimensional linear model with p exceeding n, the classical z-test (or t-test) is not applicable because it is infeasible to compute the OLS estimators of p regression coefficients. This motivates us to modify the classical z-test (or t-test) to accommodate high dimensional data.

In high dimensional regression analysis, hypothesis testing has attracted considerable attention (Goeman et al., 2006, 2011; Zhong and Chen, 2011). Since these papers mainly focus on testing a large set of coefficients against a high dimensional alternative, their approaches are not applicable for testing the significance of a single coefficient. Hence, Bühlmann (2013) recently applied the ridge estimation approach and obtained a test statistic to examine the significance of an individual coefficient. His proposed test involves a bias correction, which is different from the classical z-test (or t-test) via the OLS approach. In the meantime, Zhang and Zhang (2014) proposed a low dimensional projection procedure to construct the confidence intervals for a linear combination of a small subset of regression coefficients. The key assumption behind their procedure is the existence of good initial estimators for the unknown regression coefficients and the unknown standard deviation of random errors. To this end, the penalty function with a tuning parameter is required to implement Zhang and Zhang's (2014) procedure. Later, van de Geer et al. (2014) extended the

* Corresponding author.
E-mail addresses: lanwei@swufe.edu.cn (W. Lan), pszhong@stt.msu.edu (P.-S. Zhong), rzli@psu.edu (R. Li), hansheng@gsm.pku.edu.cn (H. Wang), cltsai@ucdavis.edu (C.-L. Tsai).

http://dx.doi.org/10.1016/j.jeconom.2016.05.016
0304-4076/

results of Zhang and Zhang's (2014) to broad models and general loss functions.

Instead of the ridge estimation and low dimensional projection, Fan and Lv (2008) and Fan et al. (2011) used the correlation approach to screen out those covariates that have weak correlations with the response variable. As a result, the total number of predictors that are highly correlated with the response variable is smaller than the sample size. However, Cho and Fryzlewicz (2012) found out that such a screening process via the marginal correlation procedure may not be reliable when the predictors are highly correlated. To this end, they proposed a tilting correlation screening (TCS) procedure to measure the contribution of the target variable to the response. Motivated by the TCS idea of Cho and Fryzlewicz (2012), we develop a new testing procedure that can lead to accurate inferences. Specifically, we adopt the TCS idea and introduce the Correlated Predictors Screening (CPS) method to control for predictors that are highly correlated with the target covariate before a hypothesis test is conducted. It is worth noting that Cho and Fryzlewicz (2012) mainly focus on variable selection, while we aim at hypothesis testing.

If the total number of highly correlated predictors resulting from the CPS procedure is smaller than the sample size, their effects can be profiled out from both the response and the target predictor via projections. Based on the profiled response and the profiled predictor, we are able to employ a classical simple regression model to obtain the OLS estimate of the target regression coefficient. We then demonstrate that the resulting estimator is $\sqrt{n}$-consistent and asymptotically normal, even if the random errors are heteroskedastic as considered by Belloni et al. (2012, 2014). Accordingly, a z-test statistic can be constructed for testing the target coefficient. Under some mild conditions, we show that the p-values obtained by the asymptotic normal distribution satisfy the weak dependence assumption of Storey et al. (2004). As a result, the multiple hypothesis testing procedure of Storey et al. (2004) can be directly applied to control the false discovery rate (FDR). Finally, we demonstrate that the proposed multiple testing procedure achieves model selection consistency.

The rest of the article is organized as follows. Section 2 introduces model notation and proposes the CPS method. The theoretical properties of hypothesis tests via the CPS as well as the FDR procedures are obtained. Section 3 presents simulation studies, while Section 4 provides real data analyses. Some concluding remarks are given in Section 5. All technical details are relegated to Appendix.

2. The methodology

2.1. The CPS method

Let (Y_i, X_i) be a random vector collected from the ith subject $(1 \le i \le n)$, where $Y_i \in \mathbb{R}^1$ is the response variable and $X_i = (X_{i1}, \ldots, X_{ip})^\top \in \mathbb{R}^p$ is the associated p-dimensional predictor vector with $E(X_i) = 0$ and $\mathrm{cov}(X_i) = \Sigma = (\sigma_{j_1 j_2}) \in \mathbb{R}^{p\times p}$. In addition, the response variable has been centralized such that $E(Y_i) = 0$. Unless explicitly stated otherwise, we hereafter assume that $p \gg n$ and n tends to infinity for asymptotic behavior. Then, consider the linear regression model,

$$Y_i = X_i^\top \beta + \varepsilon_i, \tag{2.1}$$

where $\beta = (\beta_1, \ldots, \beta_p)^\top \in \mathbb{R}^p$ is an unknown regression coefficient vector. Motivated by Belloni et al. (2012, 2014), we assume that the error terms ε_i are independently distributed with $E(\varepsilon_i|X_i) = 0$ and finite variance $\mathrm{var}(\varepsilon_i) = \sigma_i^2$ for $i = 1, \ldots, n$. In addition, define the average of error variances as $\bar\sigma_n^2 = n^{-1}\sum_i \sigma_i^2$, and assume that $\bar\sigma_n^2 \to \bar\sigma^2$ as $n \to \infty$ for some finite positive constant $\bar\sigma^2$. To assess the significance of a single coefficient, we test the null hypothesis $H_0 : \beta_j = 0$ for any given j. Without loss of generality, we focus on testing the first regression coefficient. That is,

$$H_0 : \beta_1 = 0 \quad \text{vs.} \quad H_1 : \beta_1 \neq 0, \tag{2.2}$$

and the same testing procedure is applicable to the rest of the individual regression coefficients.

For the sake of convenience, let $\mathbb{Y} = (Y_1, \ldots, Y_n)^\top \in \mathbb{R}^n$ be the vector of responses, $\mathbb{X} = (X_1, \ldots, X_n)^\top \in \mathbb{R}^{n\times p}$ be the design matrix with the jth column $\mathbb{X}_j \in \mathbb{R}^n$, and $\mathcal{E} = (\varepsilon_1, \ldots, \varepsilon_n)^\top \in \mathbb{R}^n$. In addition, let $\mathcal{A}$ be an arbitrary index set with cardinality $|\mathcal{A}|$. Then, define $X_{i\mathcal{A}} = (X_{ij} : j \in \mathcal{A})^\top \in \mathbb{R}^{|\mathcal{A}|}$, $\mathbb{X}_{\mathcal{A}} = (X_{1\mathcal{A}}, \ldots, X_{n\mathcal{A}})^\top = (\mathbb{X}_j : j \in \mathcal{A}) \in \mathbb{R}^{n\times|\mathcal{A}|}$, $\Sigma_{\mathcal{A}} = (\sigma_{j_1 j_2} : j_1 \in \mathcal{A}, j_2 \in \mathcal{A}) \in \mathbb{R}^{|\mathcal{A}|\times|\mathcal{A}|}$, and $\Sigma_{\mathcal{A}j} = \Sigma_{j\mathcal{A}}^\top = (\sigma_{j_1 j_2} : j_1 \in \mathcal{A}, j_2 = j) \in \mathbb{R}^{|\mathcal{A}|}$. Moreover, define $\Sigma_{\mathcal{A}_a\mathcal{A}_b} = (\sigma_{j_1 j_2} : j_1 \in \mathcal{A}_a, j_2 \in \mathcal{A}_b) \in \mathbb{R}^{|\mathcal{A}_a|\times|\mathcal{A}_b|}$ for any two arbitrary index sets $\mathcal{A}_a$ and $\mathcal{A}_b$, which implies $\Sigma_{\mathcal{A}\mathcal{A}} = \Sigma_{\mathcal{A}}$.

Before constructing the test statistic, we first control those predictors that are highly correlated with X_{i1}. Otherwise, they can generate a confounding effect, due to multicollinearity and yield an incorrect estimator of β_1. Specifically, the marginal regression coefficient $(\mathbb{X}_1^\top\mathbb{X}_1^\top)^{-1}\mathbb{X}_1^\top\mathbb{Y} = \beta_1 + (\mathbb{X}_1^\top\mathbb{X}_1^\top)^{-1}\mathbb{X}_1^\top(\mathbb{Y} - \mathbb{X}_1\beta_1)$ is not a consistent estimator of β_1 when $\mathbb{Y} - \mathbb{X}_1\beta_1$ and $\mathbb{X}_1$ have a strong linear relationship. To remove the confounding effect, define $\rho_{1j} = \mathrm{corr}(X_{i1}, X_{ij})$ as the correlation coefficient of X_{i1} and X_{ij} for $j = 2, \ldots, p$, and $\rho_1^* = (|\rho_{12}|, \ldots, |\rho_{1p}|)^\top \in \mathbb{R}^{p-1}$. We also assume that $|\rho_{1j}|$ are distinct. Then, let $\mathcal{S}_k$ be the set of k indices whose associated predictors have the largest absolute correlations with X_{i1}:

$$\mathcal{S}_k = \big\{2 \le j \le p : \ |\rho_{1j}| \text{ is among the first } k \text{ largest} \\ \text{absolute correlations in } \rho_1^*\big\}. \tag{2.3}$$

The choice of k (i.e., $\mathcal{S}_k$) will be discussed in Remark 2. With a slight abuse of notation, we sometimes denote $\mathcal{S}_k$ by $\mathcal{S}$ in the rest of the paper for the sake of convenience. To remove the confounding effect due to $X_{i\mathcal{S}}$, we construct the profiled response and predictor as $\widetilde{\mathbb{Y}} = \mathcal{Q}_{\mathcal{S}}\mathbb{Y}$ and $\widetilde{\mathbb{X}}_1 = \mathcal{Q}_{\mathcal{S}}\mathbb{X}_1$, respectively, where $\mathcal{Q}_{\mathcal{S}} = I_n - \mathbb{X}_{\mathcal{S}}(\mathbb{X}_{\mathcal{S}}^\top\mathbb{X}_{\mathcal{S}})^{-1}\mathbb{X}_{\mathcal{S}}^\top \in \mathbb{R}^{n\times n}$ and $I_n \in \mathbb{R}^{n\times n}$ is the $n \times n$ identity matrix. We next follow the OLS approach and obtain the estimate of the target coefficient β_1,

$$\hat\beta_1 = (\widetilde{\mathbb{X}}_1^\top\widetilde{\mathbb{X}}_1)^{-1}(\widetilde{\mathbb{X}}_1^\top\widetilde{\mathbb{Y}}) = (\mathbb{X}_1^\top\mathcal{Q}_{\mathcal{S}}\mathbb{X}_1)^{-1}(\mathbb{X}_1^\top\mathcal{Q}_{\mathcal{S}}\mathbb{Y}).$$

We refer to the above procedure as the Correlated Predictors Screening (CPS) method, $\hat\beta_1$ as the CPS estimator of β_1, and $\mathcal{S}$ as the CPS set of X_{i1}.

It is of interest to note that the proposed CPS estimator $\hat\beta_1$ is closely related to the estimator obtained via the "added-variable plot" approach (e.g., see Cook and Weisberg, 1998). To illustrate their relationship, let $\mathbb{X}_{-1}$ be the collection of all covariates in $\mathbb{X}$ except for $\mathbb{X}_1$. Then the method of "added-variable plot" essentially takes the residuals from regressing $\mathbb{Y}$ against $\mathbb{X}_{-1}$ as the response and the residuals from regressing $\mathbb{X}_1$ against $\mathbb{X}_{-1}$ as covariates. Although both approaches can be used to assess the effect of $\mathbb{X}_{-1}$ on the estimation of β_1, they are different. Specifically, the "added-variable plot" approach requires regressing $\mathbb{X}_1$ on all remaining covariates, which is not computable when the dimension p is larger than n. By contrast, CPS only considers those predictors in $\mathcal{S}$ that are highly correlated with $\mathbb{X}_1$, which is applicable in high dimensional settings.

Making inferences about β_1 in high dimensional models is challenging because these inferences can depend on the accuracy of estimating the whole vector β; see Belloni et al. (2014), van de Geer et al. (2014) and Zhang and Zhang (2014). The main contribution of our proposed CPS method is employing a simple

Please cite this article in press as: Lan, W., et al., Testing a single regression coefficient in high dimensional linear models. Journal of Econometrics (2016), http://dx.doi.org/10.1016/j.jeconom.2016.05.016