Accepted Manuscript

Sparse linear models and l_1- regularized 2SLS with high-dimensional endogenous regressors and instruments

Ying Zhu

PII: S0304-4076(17)30210-5
DOI: https://doi.org/10.1016/j.jeconom.2017.10.002
Reference: ECONOM 4433

To appear in: *Journal of Econometrics*

Received date : 23 July 2015
Revised date : 8 September 2017
Accepted date : 9 October 2017

Please cite this article as: Zhu Y., Sparse linear models and l_1- regularized 2SLS with high-dimensional endogenous regressors and instruments. *Journal of Econometrics* (2017), https://doi.org/10.1016/j.jeconom.2017.10.002

Sparse Linear Models and l_1-Regularized 2SLS with High-Dimensional Endogenous Regressors and Instruments

Ying Zhu
Department of Economics, Michigan State University
486 W. Circle Dr. Room 110, East Lansing, MI 48824. Email: yzhu@msu.edu

Abstract

We explore the validity of the 2-stage least squares estimator with l_1-regularization in both stages, for linear triangular models where the numbers of endogenous regressors in the main equation and instruments in the first-stage equations can exceed the sample size, and the regression coefficients are sufficiently sparse. For this l_1-regularized 2-stage least squares estimator, we first establish finite-sample performance bounds and then provide a simple practical method (with asymptotic guarantees) for choosing the regularization parameter. We also sketch an inference strategy built upon this practical method.

JEL Classification: C14, C31, C36
Keywords: High-dimensional statistics; Lasso; sparse linear models; endogeneity; two-stage least squares

1 Introduction

The objective of this paper is consistent estimation of regression coefficients in models with a large number of endogenous regressors and instruments. We consider the linear model

$$Y_i = X_i\beta^* + \epsilon_i = \sum_{j=1}^{p} X_{ij}\beta_j^* + \epsilon_i, \quad i = 1, ..., n \tag{1}$$

where ϵ_i is a zero-mean random error possibly correlated with X_i and β^* is a vector of unknown parameters of interest. The j^{th} component of β^* is denoted by β_j^*. The j^{th} component, X_{ij}, of the $1 \times p$ vector, X_i, is *endogenous* if $\mathbb{E}(X_{ij}\epsilon_i) \neq 0$, and *exogenous* if $\mathbb{E}(X_{ij}\epsilon_i) = 0$.

When endogenous regressors are present, the classical least squares estimator will be inconsistent for β^* (i.e., $\hat{\beta}_{OLS} \overset{p}{\nrightarrow} \beta^*$) even when the dimension p of β^* is fixed and small relative to the sample size n. The two-stage least squares (2SLS) estimation plays an important role in accounting for endogeneity that comes from individual choice or market equilibrium (e.g., Wooldridge, 2010), and is based on the following "first-stage" equations for the components of X_i,

$$X_{ij} = Z_{ij}\pi_j^* + \eta_{ij} = \sum_{l=1}^{d_j} Z_{ijl}\pi_{jl}^* + \eta_{ij}, \quad i = 1, ..., n, \; j = 1, ..., p. \tag{2}$$

For each $j = 1, ..., p$, Z_{ij} is a $1 \times d_j$ vector of instrumental variables, η_{ij} a zero-mean random error which is uncorrelated with Z_{ij}, and π_j^* is a vector of unknown nuisance parameters. We will refer to the equation in (1) as the main equation and the equations in (2) as the first-stage equations. In particular, the assumption $\mathbb{E}(Z_{ij}\epsilon_i) = \mathbb{E}(Z_{ij}\eta_{ij}) = \mathbf{0}$ for all $j = 1, ..., p$ and $\mathbb{E}(Z_{ij}\eta_{ij'}) = \mathbf{0}$ for all