Statistics and Probability Letters 169 (2021) 108967

Contents lists available at ScienceDirect

Statistics and Probability Letters

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

Endogenous treatment effect estimation using high-dimensional instruments and double selection

Wei Zhong [a,b], Yang Gao [b], Wei Zhou [a], Qingliang Fan [c,*]

[a] MOE Key Lab of Econometrics, Wang Yanan Institute for Studies in Economics, Xiamen University, China
[b] Department of Statistics and Fujian Key Lab of Statistics, School of Economics, Xiamen University, China
[c] Department of Economics, The Chinese University of Hong Kong, China

ARTICLE INFO

Article history:
Received 22 June 2020
Received in revised form 25 August 2020
Accepted 2 October 2020
Available online 14 October 2020

Keywords:
Double selection
Endogeneity
High dimensionality
Instrumental variable
Treatment effect

ABSTRACT

We propose a double selection instrumental variable estimator for the endogenous treatment effects using both high-dimensional control variables and instrumental variables. It deals with the endogeneity of the treatment variable and reduces omitted variable bias due to imperfect model selection.

1. Introduction

Estimating the causal effect of the treatment variable is of fundamental importance in the observational studies. Because subjects are usually not randomly assigned, it is necessary to assume that the treatment variable can be considered as randomly assigned after controlling for a large set of other confounding covariates (Imbens, 2004; Imbens and Rubin, 2015). But in empirical research, there is often no guidance on how to choose control variables (Donohue III and Levitt, 2001). Belloni et al. (2014) proposed a double selection (DS) method to identify the important control variables for the exogenous treatment variable. However, the treatment is often endogenous due to unavailability of important control variables or sample selection, which would lead to the inconsistency of the DS estimator. To deal with the endogeneity, the instrumental variable (IV) technique has been widely used. The optimal instrument is the conditional expectation of the endogenous variable given IVs (Amemiya, 1974). Belloni et al. (2012) proposed a post-LASSO method to select important IVs and estimate optimal instruments. Lin et al. (2015), Farrell (2015), Kang et al. (2016) and Fan and Zhong (2018) also studied high-dimensional IV models using LASSO-related methods. Zhong et al. (2020) further considered the penalized logistic regression based IV estimator for dummy treatment variable. In high dimensional data analysis literature, regularization methods have been intensively studied for variable selection, e.g., LASSO (Tibshirani, 1996), SCAD (Fan and Li, 2001), etc. However, as mentioned by Belloni et al. (2014), the traditional post-single-selection methods fail to control the omitted variables bias due to imperfect model selection. This motivates us to develop a double selection procedure for estimating the endogenous treatment effect using both high-dimensional control variables and instrumental variables.

* Correspondence to: 903, Esther Lee Building, Shatin, N.T., Hong Kong.
E-mail address: michaelqfan@gmail.com (Q. Fan).

https://doi.org/10.1016/j.spl.2020.108967
0167-7152/

In this paper, we propose a double selection instrumental variable (DS-IV) estimator using a three-step algorithm. In the first step, we select the significant control variables for the outcome using regularization methods; In the second step, we select the control variables and instrumental variables which are important to predict the endogenous variable and obtain the predicted value of the endogenous treatment variable; In the third step, we obtain the DS-IV estimator for the endogenous treatment effect based on the predicted treatment variable and the union of the selected control variables in the first two steps. The control variable selection alleviates the intrinsic difficulty of finding valid instruments. It is easily implemented using our developed R package *naivereg*[1] (Fan et al., 2020). A closely related work by Chernozhukov et al. (2015) also offers an approach to estimating structural parameters of endogenous variables in the presence of many instruments and controls. The main difference is that they used orthogonal moment functions in Belloni et al. (2014) while we focus on the selection of instrumental variables.

The rest of the paper is organized as follows: Section 2 presents the DS-IV estimator. Section 3 investigates its theoretical properties. Section 4 is a real data application. To save space, the Monte Carlo simulations and all detailed proofs are contained in the Supplementary material. Throughout the paper, we let $\|\cdot\|_0$, $\|\cdot\|$ and $\|\cdot\|_\infty$ represent the ℓ_0-norm, ℓ_2-norm and the infinity norm, respectively. $m \vee n = \max\{m, n\}$ and $\mathbb{E}_n[f] := \mathbb{E}_n[f(\omega_i)] := \sum_{i=1}^n f(\omega_i)/n$.

2. Methodology

Consider a structural equation with an endogenous treatment variable and many control variables

$$y_i = d_i\alpha_0 + \mathbf{x}_i'\boldsymbol{\beta}_0 + \varepsilon_i, \tag{2.1}$$

where y_i is the outcome variable for individual i, d_i is the endogenous treatment variable, α_0 denotes the true coefficient of d_i, $\mathbf{x}_i$ is a $p \times 1$ vector of exogenous control variables, $\boldsymbol{\beta}_0$ is a $p \times 1$ vector of the true parameters associated with $\mathbf{x}_i$, ε_i is the ith random error term for $i = 1, 2, \ldots, n$, and n is the sample size. To estimate the treatment effect accurately, we include as many as possible confounding control covariates. Thus, the dimension of $\mathbf{x}_i$, p, is allowed to be greater than n. We remark that, under the mean conditional independence assumption $\mathrm{E}[\varepsilon_i|d_i = a + 1, \mathbf{x}_i] = \mathrm{E}[\varepsilon_i|d_i = a, \mathbf{x}_i]$ for some value a in the domain of d_i, $\alpha_0 = \mathrm{E}[y_i|d_i = a + 1, \mathbf{x}_i] - \mathrm{E}[y_i|d_i = a, \mathbf{x}_i]$, which is the average treatment effect (ATE) when the treatment variable increases one unit from the value a conditional on the value of $\mathbf{x}_i$. If the treatment is binary, $\alpha_0 = \mathrm{E}[y_i|d_i = 1, \mathbf{x}_i] - \mathrm{E}[y_i|d_i = 0, \mathbf{x}_i]$ is the ATE conditional on $\mathbf{x}_i$ under the assumption $\mathrm{E}[\varepsilon_i|d_i = 1, \mathbf{x}_i] = \mathrm{E}[\varepsilon_i|d_i = 0, \mathbf{x}_i]$ which is weaker than the common conditional independence (unconfoundedness) assumption (Heckman et al., 1998).

We develop a double selection instrumental variable (DS-IV) estimator for the endogenous treatment effect α_0 with both high-dimensional control variables and instrumental variables. We denote a $q \times 1$ vector of instrumental variables by $\mathbf{w}_i = (w_{i1}, \ldots, w_{iq})'$ and consider the following linear reduced form equation of the endogenous treatment variable d_i,

$$d_i = \mathbf{w}_i'\boldsymbol{\delta}_0 + \mathbf{x}_i'\boldsymbol{\gamma}_0 + \nu_i, \tag{2.2}$$

where $\boldsymbol{\delta}_0$ and $\boldsymbol{\gamma}_0$ are $q \times 1$ and $p \times 1$ vectors of the true coefficients of instrumental variables $\mathbf{w}_i$ and the control variables $\mathbf{x}_i$ in (2.2), respectively. For notation simplicity hereafter, we rewrite Eq. (2.2) as $d_i = \mathbf{z}_i'\boldsymbol{\theta}_0 + \nu_i$, where $\mathbf{z}_i = (\mathbf{w}_i, \mathbf{x}_i)$ and $\boldsymbol{\theta}_0 = \left(\boldsymbol{\delta}_0', \boldsymbol{\gamma}_0'\right)'$. Denote by $\rho_{\varepsilon,\nu}$ the correlation between ε and ν, and $\rho_{\varepsilon,\nu} \neq 0$ indicates the endogeneity of the treatment variable d_i.

Denote the optimal instrument by $d_i^* = \mathrm{E}(d_i|\mathbf{z}_i)$. Let $\mathbf{Y} = (y_1, \ldots, y_n)'$, $\mathbf{D} = (d_1, \ldots, d_n)'$, $\mathbf{D}^* = (d_1^*, \ldots, d_n^*)'$, $\mathbf{V} = (\nu_1, \ldots, \nu_n)$ and $\widehat{\mathbf{D}} = (\widehat{d}_1^*, \ldots, \widehat{d}_n^*)'$. For $A \subset \{1, \ldots, p\}$, let $\mathbf{X}_A = \{\mathbf{X}_j, j \in A\}$ where $\mathbf{X}_j$ is the jth column vector of $\mathbf{X}$. Let $\mathcal{P}_A = \mathbf{X}_A\left(\mathbf{X}_A'\mathbf{X}_A\right)^{-1}\mathbf{X}_A'$ and $\mathcal{M}_A = \mathbf{I}_n - \mathcal{P}_A$. The DS-IV algorithm consists of the following three steps:

Step 1. We select the important control variables using regularization methods for the data $(y_i, \mathbf{x}_i)$. The objective function with ℓ_1 penalty is

$$L_{n,1}(\boldsymbol{\beta}, \lambda_n) = \sum_{i=1}^n (y_i - \mathbf{x}_i'\boldsymbol{\beta})^2 + \lambda_n \sum_{j=1}^p |\beta_j|, \tag{2.3}$$

where λ_n is a tuning parameter which is chosen by cross-validation or the BIC criterion (Wang et al., 2007). The penalized estimator $\widehat{\boldsymbol{\beta}}$ is obtained by $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_1, \widehat{\beta}_2, \ldots, \widehat{\beta}_p)' = \mathrm{argmin}_{\boldsymbol{\beta}} L_{n,1}(\boldsymbol{\beta}, \lambda_n)$. Denote $\widehat{I}_1 = \{j : \widehat{\beta}_j \neq 0, j = 1, 2, \ldots, p\}$.

Step 2 We select both important control variables and instrumental variables for the endogenous treatment in (2.2) using data $(d_i, \mathbf{w}_i, \mathbf{x}_i)$. We consider the objective function

$$L_{n,2}(\boldsymbol{\delta}, \boldsymbol{\gamma}, \lambda_n) = \sum_{i=1}^n (d_i - \mathbf{w}_i'\boldsymbol{\delta} - \mathbf{x}_i'\boldsymbol{\gamma})^2 + \lambda_n\Big(\sum_{j=1}^q |\delta_j| + \sum_{j=1}^p |\gamma_j|\Big), \tag{2.4}$$

where the estimator $\widehat{\boldsymbol{\delta}}$ and $\widehat{\boldsymbol{\gamma}}$ are obtained by minimizing the objective function $L_{n,2}(\boldsymbol{\delta}, \boldsymbol{\gamma}, \lambda_n)$ in (2.4), i.e. $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\delta}}', \widehat{\boldsymbol{\gamma}}')' = \mathrm{argmin}_{\boldsymbol{\delta},\boldsymbol{\gamma}} L_{n,2}(\boldsymbol{\delta}, \boldsymbol{\gamma}, \lambda_n)$. Denote $\widehat{I}_2 = \{j : \widehat{\gamma}_j \neq 0, j = 1, 2, \ldots, p\}$. Then, the optimal instrument is estimated by $\widehat{d}_i^* = \mathbf{w}_i'\widehat{\boldsymbol{\delta}} + \mathbf{x}_i'\widehat{\boldsymbol{\gamma}} = \mathbf{z}_i'\widehat{\boldsymbol{\theta}}$.

[1] https://cran.r-project.org/web/packages/naivereg/index.html.

Step 3 We define the DS-IV estimator $\widehat{\alpha}$ for the endogenous treatment effect based on the predicted treatment variable $\widehat{d}_i^*$ and the union of selected control variables in the first two variable selection steps denoted by $\widehat{I} = \widehat{I}_1 \cup \widehat{I}_2$. That is,

$$\widehat{\alpha} = \left(\widehat{\mathbf{D}}' \mathcal{M}_{\widehat{I}} \mathbf{D}\right)^{-1} \left(\widehat{\mathbf{D}}' \mathcal{M}_{\widehat{I}} \mathbf{Y}\right). \tag{2.5}$$

Remark. The double selection is robust to imperfect model selection of single selection methods. Some key control variables might be missed by the first-step single selection using the data $(y_i, \mathbf{x}_i)$ because either the beta-min condition cannot be satisfied (Van De Geer et al., 2011) or the sample size is limited. If these missed key controls happen to be correlated with the treatment variable, then the treatment variable is endogenous due to the omitted key variables which lead to estimation bias. Thus, the second step is crucial to ensure the accuracy of the estimated treatment effect. We also remark that this algorithm can be easily extended to the multi-dimensional endogenous variables using the similar framework.

3. Theoretical properties

Assume $\|\boldsymbol{\beta}_0\|_0 \leq s$ and $\|\boldsymbol{\gamma}_0\|_0 \leq s$, where the number of true regression coefficients $\boldsymbol{\beta}_0$ and $\boldsymbol{\gamma}_0$ cannot exceed $s \ll n$, and its estimator is given by $\widehat{s} = \|\widehat{I}\|_0$. The following regularity conditions are imposed for the theoretical properties of the proposed DS-IV estimator.

(A) Define $\phi_{\min}(m)[M] := \min_{1\leq\|\delta\|_0\leq m} \frac{\delta' M \delta}{\|\delta\|^2}$ and $\phi_{\max}(m)[M] := \max_{1\leq\|\delta\|_0\leq m} \frac{\delta' M \delta}{\|\delta\|^2}$ for a semi-definite matrix M. There is an absolute sequence $a_n \to \infty$ such that with probability at least $1 - \Delta_n$, $k' \leq \phi_{\min}(a_n s)[\mathbb{E}_n(\mathbf{z}_i \mathbf{z}_i')] \leq \phi_{\max}(a_n s)[\mathbb{E}_n(\mathbf{z}_i \mathbf{z}_i')] \leq k''$, where $\mathbb{E}_n(\mathbf{z}_i \mathbf{z}_i') = \sum_{i=1}^n \mathbf{z}_i \mathbf{z}_i'/n$ and $0 < k' < k'' < \infty$ are absolute constants.
(B) For $i = 1, \ldots, n$, the two error terms satisfy $\mathrm{E}(\varepsilon_i^2) < \infty$ and $\mathrm{E}(v_i^2) < \infty$.
(C) $\log m = o(n^{1/3})$ and $s \log(m \vee n)/n \to 0$, where $m = p + q$. There exists a constant C such that $E(x_{ij}^3 v_i^3) \leq C$.
(D) Assume that $\mathrm{E}(d_i^{*2}) < \infty$ and $\max_{1\leq j\leq s} |\sum_{i=1}^n x_{ij} d_i^*/n| < \infty$ hold.

Condition (A) and Condition (C) are extended from Condition SE(P) for Sparse Eigenvalues in Belloni et al. (2014). Condition (A) can directly hold for $\mathbf{x}_i, i = 1, \ldots, n$ being i.i.d. zero-mean sub-Gaussian random vectors or i.i.d. bounded zero-mean random vectors. Condition (C) also guarantees the validity of the proposed method can deal with the high dimensionality of the control variables and IVs, also can be applied to moderate deviation theorems for self-normalized sums to obtain a bound for some error components. Condition (B) is the moment conditions imposed on the error terms. Condition (D) imposes mild restriction on the moment of some important term involving the endogenous treatment variable.

Lemma 3.1 (*Model Selection Consistency*). *Under Conditions (A), (B) and (C), we have* $\|\widehat{\mathbf{D}} - \mathbf{D}^*\| = o_p\left(\sqrt{s \log(m \vee n)/n}\right)$, $\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_1 = o_p\left(s\sqrt{\log(m \vee n)/n}\right)$, $\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_2 = o_p\left(\sqrt{s \log(m \vee n)/n}\right)$.

Lemma 3.1 gives the consistency for (Post-)LASSO estimators, which can be derived from Lemma 1 in Belloni et al. (2014). The main result of the DS-IV estimator in Theorem 3.1.

Theorem 3.1. *Suppose Conditions (A)–(D) hold, the DS-IV estimator* $\widehat{\alpha}$ *of the endogenous treatment effect* α_0 *is root-n consistent and asymptotically normal. That is,* $\sigma_n^{-1}\sqrt{n}(\widehat{\alpha} - \alpha_0) \to N(0, 1)$ *in distribution, where* $\sigma_n^2 = \left(\mathrm{E}(\mathbf{D}^{*\prime} \mathcal{M}_I \mathbf{D}^*)\right)^{-1} \mathrm{E}\left(\mathbf{D}^{*\prime} \mathcal{M}_I \mathbf{D}^* \varepsilon_i^2\right) \left(\mathrm{E}(\mathbf{D}^{*\prime} \mathcal{M}_I \mathbf{D}^*)\right)^{-1}$. *If* $\mathrm{E}(\varepsilon_i^2) = \sigma_\varepsilon^2$ *almost surely for all* $1 \leq i \leq n$, *then* $\sigma_n^2 = \left(\mathrm{E}(\mathbf{D}^{*\prime} \mathcal{M}_I \mathbf{D}^*)\right)^{-1} \sigma_\varepsilon^2$.

Theorem 3.1 demonstrates that the DS-IV estimator is asymptotically unbiased, root-n consistent and asymptotically normal. It is parallel with Theorem 3.1 in Belloni et al. (2014). However, it is different from Belloni et al. (2014) in that our DS-IV estimator is able to deal with the endogeneity of the treatment variable.

4. The treatment effect of teacher's attentiveness on student's achievement

Home visits could facilitate parent involvement, reduce discipline problems and increase student's overall positive attitudes toward school (Dohl and Lochner, 2012; Castro et al., 2015). Using a comprehensive survey data, the China Education Panel Survey (CEPS), we investigate the treatment effects of home visits on students' performance measured by standardized exam grades. We define the treatment variable d_i to be 1 if the class adviser goes to the ith student's home to talk with the parents at least once during the school year and 0 otherwise. The response variable $score_i$ is one of Mathematics, Chinese, English subject score of the ith student.

We consider

$$score_i = d_i \alpha_0 + \mathbf{x}_i' \boldsymbol{\beta}_0 + \varepsilon_i,$$

Table 1
Estimated effects of home visit on school performance.

	Math Score		Chinese Score		English Score	
	Effect	Std. Err.	Effect	Std. Err.	Effect	Std. Err.
OLS (univariate)	−0.150	0.198	−0.159	0.198	−0.379*	0.198
OLS (multivariate)	0.003	0.212	0.052	0.199	−0.055	0.201
TSLS	0.624**	0.297	0.556*	0.290	0.497*	0.292
DS	0.337*	0.201	0.381*	0.210	0.225	0.204
Post-LASSO	0.730**	0.317	0.646*	0.355	0.538*	0.322
DS-IV	1.273***	0.291	1.159***	0.304	1.218***	0.310

Notes: OLS (univariate) is the univariate OLS with only treatment variable. OLS (multivariate) is multivariate regression.
*Denotes 0.1 significance level.
**Denotes 0.05 significance level.
***Denotes 0.01 significance level.

where $\mathbf{x}_i$ is a set of covariates which include student's gender, cognitive ability test score, health status, total hours of study after school, mother's education, family income, number of siblings, whether student lives with grand parents, and the characteristics of the response variable subject teacher including age, gender, total teaching hours, etc. We employ the IV method to address the endogeneity issue of the home visit d_i due to unobserved common factors such as student's true interests in studying that subject. The candidate IVs include the age, marital status, gender, the number of other classes, total teaching hours last week, total teaching preparation hours last week, the number of hours spent on grading homework of the last week of the non-adviser teachers of a different subject from the response. We also include the polynomial terms up to order 3 and the interactions of all these variables. For potential covariates, we also include the student height, weight, ethnicity, previous year grades, sleeping time, extra curriculum activities, family activities, etc. The sample size is 7617. The percentage of observations with home visit is 42.93%.

The regression results are presented in Table 1. First, we see that the ordinary least square (OLS) estimation in general reports no significant effect of home visits on any of three subjects, even negative effect for English scores. This is because the OLS estimator is severely biased due to the endogeneity. The two-stage least square method (TSLS) shows home visits can improve the standardized exam grades by around 0.5 points (in the 100-point scale) on average, holding other factors constant. The DS (Belloni et al., 2014) is supposed to select the important covariates of student performance, but it might still be biased due to the endogeneity of the treatment. The magnitude of treatment effects using the DS is about half the size of TSLS. The post-LASSO (Belloni et al., 2012) results show a slightly larger treatment effect than the TSLS. The proposed DS-IV method shows the strongest evidence of home visits on students performance in all three subjects. Home visits could significantly improve on average 1.2 points of standardized grades.

Acknowledgments

The authors thank Laura M. Sangalli (the editor), the associate editor, and the anonymous referees for their helpful comments that improved the article significantly. Zhong's work is supported by National Natural Science Foundation of China (11671334 and 11922117) and Fujian Provincial Natural Science Fund for Distinguish Young Scholars (2019J06004). Fan's research, in part, was supported by the National Natural Science Foundation of China Grants 71671149, 71801183 and 71631004 (Key Project) and the Science Foundation of the Ministry of Education of China (18YJC790073).

Appendix A. Supplementary materials

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.spl.2020.108967. All technical proofs and Monte Carlo simulations are included in the online Supplemental material.

References

Amemiya, T., 1974. The non-linear two-stage least squares estimator. J. Econometrics 2, 105–110.
Belloni, A., Chen, D., Chernozhukov, V., Hansen, C., 2012. Sparse models and methods for optimal instruments with an application to eminent domain. Econometrica 80, 2369–2429.
Belloni, A., Chernozhukov, V., Hansen, C., 2014. Inference on treatment effects after selection amongst high-dimensional controls. Rev. Econom. Stud. 81, 608–650.
Castro, M., Expósito-Casas, E., López-Martín, E., Lizasoain, L., Navarro-Asencio, E., Gaviria, J.L., 2015. Parental involvement on student academic achievement: A meta-analysis. Educ. Res. Rev. 14, 33–46.
Chernozhukov, V., Hansen, C., Spindler, M., 2015. Post-selection and post-regularization inference in linear models with many controls and instruments. Amer. Econ. Rev.: Pap. Proc. 105, 486–490.
Dohl, G., Lochner, L., 2012. The impact of family income on child achievement: Evidence from the earned income tax credit. Amer. Econ. Rev. 102, 1927–1956.
Donohue III, J.J., Levitt, S.D., 2001. The impact of legalized abortion on crime. Q. J. Econ. 116, 379–420.

Fan, Q., He, K., Zhong, W., 2020. R package *naivereg*: Nonparametric additive instrumental variable estimator and related IV methods. Version 1.0.5. Published on 2020-03-18. https://cran.r-project.org/web/packages/naivereg/index.html.
Fan, J., Li, R., 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. J. Amer. Statist. Assoc. 96, 1348–1360.
Fan, Q., Zhong, W., 2018. Nonparametric additive instrumental variable estimator: A group shrinkage estimation perspective. J. Bus. Econom. Statist. 36, 388–399.
Farrell, M.H., 2015. Robust inference on average treatment effects with possibly more covariates than observations. J. Econometrics 189, 1–23.
Heckman, J., Ichimura, H., Smith, J., Todd, P., 1998. Characterizing selection bias using experimental data. Econometrica 66, 1017–1098.
Imbens, G.W., 2004. Nonparametric estimation of average treatment effects under exogeneity: A review. Rev. Econ. Stat. 86, 4–29.
Imbens, G.W., Rubin, D.B., 2015. Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction. Cambridge University Press, New York.
Kang, H., Zhang, A., Cai, T.T., Small, D.S., 2016. Instrumental variables estimation with some invalid instruments and its application to mendelian randomization. J. Amer. Statist. Assoc. 111, 132–144.
Lin, W., Feng, R., Li, H., 2015. Regularization methods for high-dimensional instrumental variables regression with an application to genetical genomics. J. Amer. Statist. Assoc. 110, 270–288.
Tibshirani, R., 1996. Regression shrinkage and selection via lasso. J. R. Statist. Soc. Ser. B-Statist. Methodol. 58, 267–288.
Van De Geer, S., Bühlmann, P., Zhou, S., 2011. The adaptive and the thresholded lasso for potentially misspecified models. Electron. J. Stat. 5, 688–749.
Wang, H., Li, R., Tsai, C., 2007. Tuning parameter selectors for the smoothly clipped absolute deviation method. Biometrika 94, 553–568.
Zhong, W., Zhou, W., Fan, Q., Gao, Y., 2020. Dummy Endogenous Treatment Effect Estimation using High-Dimensional Instrumental Variables. Working paper.