



AgEcon SEARCH
RESEARCH IN AGRICULTURAL & APPLIED ECONOMICS

The World's Largest Open Access Agricultural & Applied Economics Digital Library

This document is discoverable and free to researchers across the globe due to the work of AgEcon Search.

Help ensure our sustainability.

Give to AgEcon Search

AgEcon Search
<http://ageconsearch.umn.edu>
aesearch@umn.edu

*Papers downloaded from **AgEcon Search** may be used for non-commercial purposes and personal study only. No other use, including posting to another Internet site, is permitted without permission from the copyright owner (not AgEcon Search), or as allowed under the provisions of Fair Use, U.S. Copyright Act, Title 17 U.S.C.*



csa2sls: A complete subset approach for many instruments using Stata

Seojeong Lee
Department of Economics
Seoul National University
Seoul, Korea
s.jay.lee@snu.ac.kr

Siha Lee
Department of Economics
McMaster University
Hamilton, Canada
lees223@mcmaster.ca

Julius Owusu
Department of Economics
McMaster University
Hamilton, Canada
owusuj4@mcmaster.ca

Youngki Shin
Department of Economics
McMaster University
Hamilton, Canada
shiny11@mcmaster.ca

Abstract. We developed a command, `csa2sls`, that implements the complete subset averaging two-stage least-squares (CSA2SLS) estimator in Lee and Shin (2021, *Econometrics Journal* 24: 290–314). The CSA2SLS estimator is an alternative to the two-stage least-squares estimator that remedies the bias issue caused by many correlated instruments. We conduct Monte Carlo simulations and confirm that the CSA2SLS estimator reduces both the mean squared error and the estimation bias substantially when instruments are correlated. We illustrate the usage of `csa2sls` in Stata with an empirical application.

Keywords: `st0732`, `csa2sls`, many instruments, complete subset averaging, two-stage least squares

1 Introduction

The two-stage least-squares (2SLS) estimator is one of the most widely used methods in applied economics. Theoretically, the optimal instrument can be achieved by the conditional mean function of the first-stage regression. However, in practice, practitioners working with a finite sample face a crucial question of how many instruments one should use, especially when there are many instruments available. This is partly due to the well-known tradeoff between bias and variance when the number of instruments increases. Donald and Newey (2001) show this point clearly with a higher-order Nagar expansion and propose choosing the optimal number of instruments that minimizes the mean squared errors (MSEs). Kuersteiner and Okui (2010) propose a model averaging approach for the first-stage regression and show that it achieves the optimal weight. These other approaches, however, require the practitioner to either know the order of importance among instruments (Donald and Newey 2001) because the method chooses the first few important instruments or estimate the optimal weights for the instruments (Kuersteiner and Okui 2010).

As an alternative, Lee and Shin (2021) propose a model-averaging approach that uses all size- k subsets of the set of available instruments in a cross-sectional regression model. This new approach is named the complete subset averaging two-stage least-squares (CSA2SLS) estimator. One advantage of the CSA2SLS estimator is that, because it uses all subsets, it does not require knowledge of the order of importance among instruments. Furthermore, averaging models using equal weights reduces potential efficiency loss in finite samples. This is because when estimated weights (instead of equal weights) are used, these become additional parameters in the model and therefore cause inefficiency when there are many models to be averaged.

We developed a command, `csa2s1s`, that implements the CSA2SLS estimator. It selects the optimal number of subset size k that minimizes the approximate MSEs. Because the size of the complete subset grows at the order of 2^K , where K is the total number of instruments, CSA2SLS is computationally intensive. To alleviate such a computational burden, the command `csa2s1s` includes options for subsampling and a fast but memory-intensive method.

The remainder of this article is organized as follows. Section 2 introduces the CSA2SLS estimator in Lee and Shin (2021). Section 3 explains the command `csa2s1s`. Section 4 shows results from Monte Carlo experiments that numerically illustrate how the CSA2SLS estimator alleviates some of the issues that arise from many instruments. Section 5 provides an empirical application of `csa2s1s`. Section 6 concludes.

2 CSA2SLS estimator

In this section, we explain the key idea of the CSA2SLS estimator in Lee and Shin (2021). Heuristically speaking, we estimate the first-stage predicted value by model averaging and apply the 2SLS estimation with those predicted values. Given a total of K instruments, we consider all subsets composed of k instruments. We compute a simple average of predicted values across models, and the 2SLS estimator follows immediately. The optimal k is selected by minimizing the approximate MSEs criterion, which will be explained in detail below.

To be concrete, consider the following model generated from an independent and identically distributed sample:

$$y_i = \mathbf{Y}_i' \boldsymbol{\beta}_y + \mathbf{x}_{1i}' \boldsymbol{\beta}_x + \epsilon_i = \mathbf{X}_i' \boldsymbol{\beta} + \epsilon_i$$

$$\mathbf{X}_i = \begin{bmatrix} \mathbf{Y}_i \\ \mathbf{x}_{1i} \end{bmatrix} = \mathbf{f}(\mathbf{z}_i) + \mathbf{u}_i = \begin{bmatrix} E(Y_i | \mathbf{z}_i) \\ \mathbf{x}_{1i} \end{bmatrix} + \begin{bmatrix} \boldsymbol{\eta}_i \\ \mathbf{0} \end{bmatrix}, \quad i = 1, \dots, N$$

where y_i is a scalar outcome variable, \mathbf{Y}_i is a $d_1 \times 1$ vector of endogenous variables, \mathbf{x}_{1i} is a $d_2 \times 1$ vector of included exogenous variables, \mathbf{z}_i is a vector of exogenous variables (including \mathbf{x}_{1i}), $\mathbf{f}(\cdot)$ is an unknown function of \mathbf{z} , and ϵ_i and \mathbf{u}_i are error terms uncorrelated with \mathbf{z}_i . Finally, $\boldsymbol{\eta}_i$ denotes an error term when we project the endogenous regressor \mathbf{Y}_i into the space of exogenous variable \mathbf{z}_i . Note that $E(\boldsymbol{\eta}_i | \mathbf{z}_i) = \mathbf{0}$ by construction.

Let $\mathbf{y} = (y_1, \dots, y_N)'$, $\boldsymbol{\epsilon} = (\epsilon_1, \dots, \epsilon_N)'$, $\mathbf{X} = (\mathbf{X}_1, \dots, \mathbf{X}_N)'$, $\mathbf{f} = (\mathbf{f}_1, \dots, \mathbf{f}_N)'$, and $\mathbf{U} = (\mathbf{u}_1, \dots, \mathbf{u}_N)'$, where $\mathbf{f}_i = \mathbf{f}(\mathbf{z}_i)$. The set of instruments has the form $\mathbf{Z}_{K,i} \equiv \{\psi_1(\mathbf{z}_i), \dots, \psi_K(\mathbf{z}_i), \mathbf{x}_{1i}\}'$, where ψ_k 's are functions of \mathbf{z}_i such that $\mathbf{Z}_{K,i}$ is the collection of $(K + d_2)$ instruments. Note that the total number of instruments K can increase as $N \rightarrow \infty$. We suppress the dependency of K on N for notation simplicity. Let $\mathbf{Z}_K = (\mathbf{Z}_{K,1}, \dots, \mathbf{Z}_{K,N})'$ be the collection of $\mathbf{Z}_{K,i}$.

Let M be the number of subsets (or models) with k instruments:

$$M = \binom{K}{k} = \frac{K!}{k!(K-k)!}$$

We also suppress the dependency of M on K and k . Let $m \in \{1, \dots, M\}$ be an index of each model and $\mathbf{z}_{m,i}^k$ be a vector of instruments in model m . Then the first-stage regression of model m can be written as

$$\mathbf{X} = \mathbf{\Pi}_m^{k'} \mathbf{Z}_m^k + \mathbf{u}_m^k$$

The average predicted value of \mathbf{X} is

$$\hat{\mathbf{X}} = \frac{1}{M} \sum_{m=1}^M \mathbf{Z}_m^k \hat{\mathbf{\Pi}}_m^k$$

where $\hat{\mathbf{\Pi}}_m^k$ is the ordinary least-squares (OLS) estimator of $\mathbf{\Pi}_m^k$. Then the CSA2SLS estimator is defined as

$$\hat{\boldsymbol{\beta}} = (\hat{\mathbf{X}}' \mathbf{X})^{-1} \hat{\mathbf{X}}' \mathbf{y}$$

Using the projection matrices, we can also write the CSA2SLS estimator as a one-step procedure,

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}' \mathbf{P}^k \mathbf{X})^{-1} \mathbf{X}' \mathbf{P}^k \mathbf{y}$$

where $\mathbf{P}^k = M^{-1} \sum_{m=1}^M \mathbf{P}_m^k$ with $\mathbf{P}_m^k = \mathbf{Z}_m^k (\mathbf{Z}_m^{k'} \mathbf{Z}_m^k)^{-1} \mathbf{Z}_m^{k'}$.

The optimal subset size k is chosen by minimizing the approximate MSE. Let $\tilde{\boldsymbol{\beta}}$ be a preliminary estimator and $\tilde{\boldsymbol{\epsilon}} = \mathbf{y} - \mathbf{X} \tilde{\boldsymbol{\beta}}$. The fitted value of \mathbf{f} is given as

$$\tilde{\mathbf{f}} = \tilde{\mathbf{Z}}^k (\tilde{\mathbf{Z}}^{k'} \tilde{\mathbf{Z}}^k)^{-1} \tilde{\mathbf{Z}}^{k'} \mathbf{X}$$

where $\tilde{\mathbf{Z}}^k$ consists of exogenous variables plus the preliminary selection of instruments as described above. Let $\tilde{\mathbf{P}}_{\tilde{\mathbf{Z}}} = \tilde{\mathbf{Z}}^k (\tilde{\mathbf{Z}}^{k'} \tilde{\mathbf{Z}}^k)^{-1} \tilde{\mathbf{Z}}^{k'}$. The residual matrix is denoted by $\tilde{\mathbf{u}} = \mathbf{X} - \tilde{\mathbf{f}}$. Define $\tilde{\mathbf{H}} = \tilde{\mathbf{f}}' \tilde{\mathbf{f}} / N$, $\tilde{\sigma}_{\epsilon}^2 = \tilde{\boldsymbol{\epsilon}}' \tilde{\boldsymbol{\epsilon}} / N$, $\tilde{\sigma}_{u\epsilon} = \tilde{\mathbf{u}}' \tilde{\boldsymbol{\epsilon}} / N$, $\tilde{\sigma}_{\lambda\epsilon} = \tilde{\boldsymbol{\lambda}}' \tilde{\mathbf{H}}^{-1} \tilde{\sigma}_{u\epsilon}$, and $\tilde{\Sigma}_u = \tilde{\mathbf{u}}' \tilde{\mathbf{u}} / N$. Then the sample counterpart of the approximate MSE is given by

$$\hat{S}_{\lambda}(k) = \tilde{\sigma}_{\lambda\epsilon}^2 \frac{k^2}{N} + \tilde{\sigma}_{\epsilon}^2 \left(\tilde{\boldsymbol{\lambda}}' \tilde{\mathbf{H}}^{-1} \tilde{\mathbf{e}}_f^k \tilde{\mathbf{H}}^{-1} \tilde{\boldsymbol{\lambda}} - \tilde{\boldsymbol{\lambda}}' \tilde{\mathbf{H}}^{-1} \tilde{\boldsymbol{\xi}}_f^k \tilde{\mathbf{H}}^{-1} \tilde{\boldsymbol{\xi}}_f^k \tilde{\mathbf{H}}^{-1} \tilde{\boldsymbol{\lambda}} \right)$$

where

$$\begin{aligned}\tilde{\mathbf{e}}_f^k &= \frac{\mathbf{X}'(\mathbf{I} - \mathbf{P}_k)^2 \mathbf{X}}{N} + \tilde{\Sigma}_u \left[\frac{2k - \text{tr}\{(\mathbf{P}^k)^2\}}{N} \right] \\ \tilde{\xi}_f^k &= \frac{\mathbf{X}'(\mathbf{I} - \mathbf{P}_k)^2 \mathbf{X}}{N} + \tilde{\Sigma}_u \frac{k}{N} - \tilde{\Sigma}_u \\ \tilde{\sigma}_{\lambda\epsilon}^2 &= \left(\tilde{\lambda}' \tilde{\mathbf{H}}^{-1} \tilde{\sigma}_{\lambda\epsilon} \right)^2\end{aligned}$$

The preliminary estimator $\tilde{\beta}$ can be estimated either by using Mallows's two-step criterion or by adopting the one-step method. See Lee and Shin (2021) for details.

3 The `csa2sls` command

3.1 Syntax

The syntax for the command is as follows:

```
csa2sls depvar [varlist1] (varlist2 = varlist_iv) [if] [in] [, noconstant
    hasconstant onestep r(#) vce(vctype) level(#) first small large
    noheader depname(depname) perfect]
```

varlist1 is the list of exogenous variables. *varlist2* is the list of endogenous variables. *varlist_iv* is the list of exogenous variables used with *varlist1* as instruments for *varlist2*.

3.2 Options

`noconstant`; see [R] **Estimation options**.

`hasconstant` indicates that a user-defined constant or its equivalent is specified among the independent variables.

`onestep` allows the one-step preliminary method. The default is Mallows's two-step criterion. See Lee and Shin (2021).

`r(#)` specifies a positive integer for the maximum number of randomly selected subsets when the number of subsets is bigger than `#`. This is useful because the number of subsets depends exponentially on the number of instruments.

`vce(vctype)` specifies the type of standard error reported, which includes types that are robust to some kinds of misspecification (`robust`) and that allow for intragroup correlation (`cluster clustvar`). `vce(unadjusted)` specifies that an unadjusted (non-robust) variance-covariance estimate matrix be used.

`level(#)`; see [R] **Estimation options**.

first requests that the first-stage regression results be displayed.

small requests that the degrees-of-freedom adjustment $N/(N - k)$ be made to the variance-covariance matrix of parameters and that small-sample F and t statistics be reported, where N is the sample size and k is the number of parameters estimated. By default, no degrees-of-freedom adjustment is made, and Wald and z statistics are reported. Even with this option, no degrees-of-freedom adjustment is made to the weighting matrix when the generalized method of moments estimator is used.

large turns on the large-sample estimation program. When the sample size is large, the average projection matrices may require a large memory size. The large option must be turned on to avoid an insufficient memory issue. The default is not using this option.

noheader suppresses the display of the summary statistics at the top of the output, displaying only the coefficient table.

depname(*depname*) specifies to substitute the dependent variable name.

perfect requests that **csa2sls** not check for collinearity between the endogenous regressors and excluded instruments, allowing one to specify “perfect” instruments. This option may be required when using **csa2sls** to implement other estimators.

3.3 Stored results

csa2sls stores the following in **e()**:

Scalars

e(N)	number of observations
e(df_m)	model degrees of freedom
e(chi2)	χ^2
e(rank)	rank of e(V)
e(rss)	residual sum of squares
e(optimal_k)	optimal subset size of instruments
e(rmse)	root MSE
e(mss)	model sum of squares
e(r2)	R^2
e(r2_a)	adjusted R^2

Macros

e(cmd)	csa2sls
e(cmdline)	command as typed
e(depvar)	name of the dependent variable
e(title)	title in estimation output
e(clustvar)	name of cluster variable
e(properties)	b V
e(predict)	program used to implement predict
e(footnote)	program used to implement footnote display
e(exogr)	name of the exogenous variables
e(insts)	name of the instruments
e(instd)	name of the instrumented variables
e(constant)	noconstant or hasconstant if specified
e(Premethod)	Mallows Criterion or One Step

Matrices	
$\mathbf{e}(\mathbf{b})$	coefficient matrix
$\mathbf{e}(\mathbf{V})$	variance-covariance matrix
Functions	
$\mathbf{e}(\text{sample})$	marks estimation sample

4 Monte Carlo experiments

In this section, we conduct Monte Carlo simulation studies focusing on the effect of correlated instruments. An independent and identically distributed sample (y_i, Y_i, \mathbf{z}_i) is generated from the following simulation design:

$$\begin{aligned} y_i &= \beta_0 + \beta_1 Y_i + \epsilon_i \\ Y_i &= \boldsymbol{\pi}' \mathbf{z}_i + u_i \end{aligned}$$

where Y_i is a scalar endogenous regressor, (β_0, β_1) is set to be $(0, 0.1)$, and \mathbf{z}_i is a K -dimensional vector of instruments generated from a multivariate normal distribution $N(\mathbf{0}, \boldsymbol{\Sigma}_z)$. The diagonal elements of $\boldsymbol{\Sigma}_z$ are set to be 1, and the off-diagonal elements are ρ_z . We set each element of $\boldsymbol{\pi}$ to be $\sqrt{0.1 / \{K + K(K-1)\rho_z(1-0.1)\}}$, where 0.1 is the R^2 in the first-stage regression. The vector of error terms (ϵ_i, u_i) follows a bivariate normal distribution whose means are zeros and variances are ones. The covariance between ϵ_i and u_i is set to be 0.9. In these simulation studies, K varies in $\{5, 10, 15, 20\}$ and ρ_z varies in $\{0, 0.5, 0.9\}$. The sample size is set to be $n = 100$, and the results are from 1,000 replications.

Figure 1 summarizes the simulation results. We report the mean bias and MSE of CSA2SLS along with the performance of the OLS estimator and the 2SLS estimator. First, the CSA2SLS estimator reduces the bias substantially when instruments are correlated ($\rho_z = 0.5, 0.9$). As predicted by theory, the bias of 2SLS increases as K increases. Note that when instruments are independent ($\rho_z = 0.0$), the difference in the bias between the CSA2SLS estimator and the 2SLS estimator is small. Lee and Shin (2021) prove that the performance of CSA2SLS will be asymptotically equivalent to that of 2SLS when $\rho_z = 0$.

Second, the efficiency loss of CSA2SLS is modest. When instruments are correlated, CSA2SLS achieves lower MSEs when $K \geq 10$. Like the bias, the MSE gap between CSA2SLS and 2SLS increases as K increases. It is also worthwhile to note that the MSE of CSA2SLS does not change much over different values of K . Finally, the OLS estimator performs the worst in these simulation designs.

To summarize, the CSA2SLS estimator shows a good finite sample performance as predicted by theory. We also observe the increased bias of 2SLS when there are many instruments. We recommend practitioners use the CSA2SLS estimator when they have many *correlated* instruments.

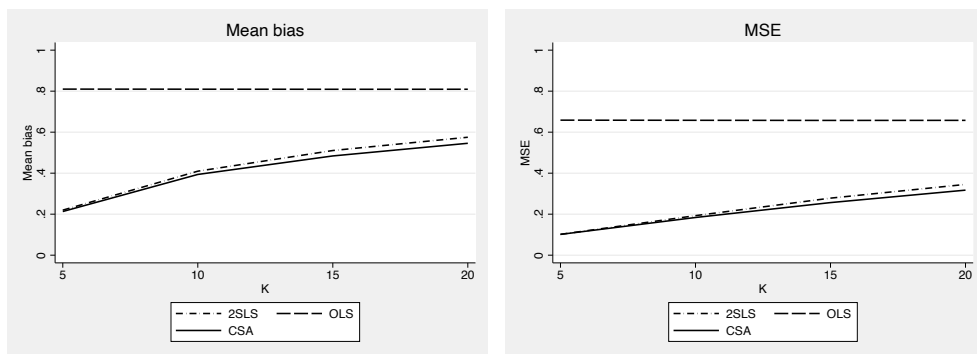
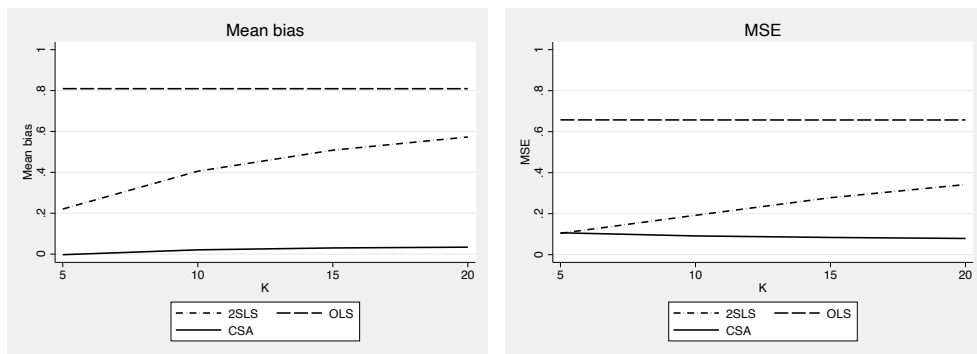
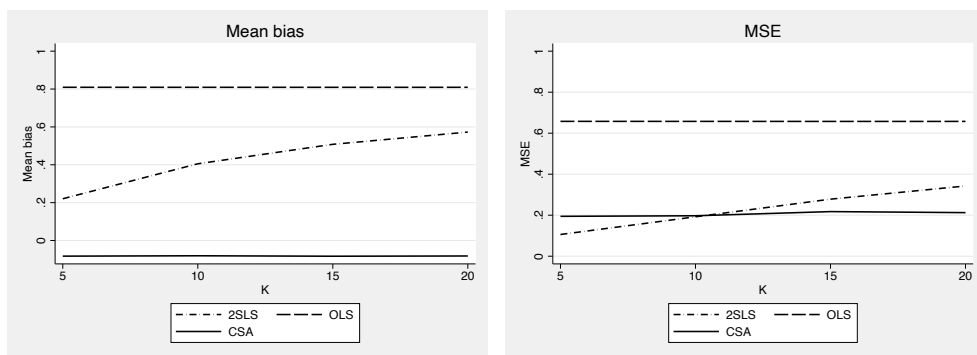
(a) $\rho_z = 0.0$ (b) $\rho_z = 0.5$ (c) $\rho_z = 0.9$

Figure 1. Mean bias and MSE

5 Empirical illustration

In this section, we illustrate the usage of `csa2s1s` with an empirical application. In this example, we revisit Berry, Levinsohn, and Pakes (1995) and estimate a logistic demand function for automobiles based on pooled cross-sectional data over different markets.

The model is specified as

$$\begin{aligned}\log(S_i) - \log(S_0) &= \alpha_0 P_i + \mathbf{X}_i' \beta_0 + \epsilon_i \\ P_i &= \mathbf{Z}_i' \delta_0 + \mathbf{X}_i' \rho_0 + u_i\end{aligned}$$

where S_i is the market share of product i with product 0 denoting the outside option, P_i is the endogenous price variable, \mathbf{X}_i is a vector of included exogenous variables, and \mathbf{Z}_i is a set of 10 instruments. The parameter of interest is α_0 , from which we can calculate the price elasticity of demand. Note that the optimal subset size k is 9 in this empirical example.

```
. set seed 2022
. insheet using blp.csv, comma
(54 vars, 2,217 obs)

. csa2s1s y hpwt air mpd space (price = sumother1 sumotherhpwt sumotherair
> sumothermpd sumotherspace sumrival1 sumrivalhpwt sumrivalair sumrivalmpd
> sumrivalspace)

Complete Subset Model Averaging 2SLS Regression      Number of obs   =      2,217
                                                    Wald chi2(5)    =      820.64
                                                    Prob > chi2     =      0.0000
                                                    R-squared       =      0.3373
                                                    Root MSE      =      1.1245
```

y	Coefficient	Std. err.	z	P> z	[95% conf. interval]	
price	-.142563	.0117095	-12.18	0.000	-.1655131	-.1196128
hpwt	1.422452	.414676	3.43	0.001	.6097024	2.235202
air	.5620958	.1379201	4.08	0.000	.2917772	.8324143
mpd	.1579617	.0471821	3.35	0.001	.0654864	.2504369
space	2.284253	.1289588	17.71	0.000	2.031499	2.537008
_cons	-2.342198	.2673599	-8.76	0.000	-2.866214	-1.818182

```
Endogenous: price
Exogenous:  hpwt air mpd space sumother1 sumotherhpwt sumotherair
             sumothermpd sumotherspace sumrival1 sumrivalhpwt sumrivalair
             sumrivalmpd sumrivalspace
```

```
. correlate sumother1 sumotherhpwt sumotherair sumothermpd sumotherspace
> sumrival1 sumrivalhpwt sumrivalair sumrivalmpd sumrivalspace
(obs=2,217)
```

	sumoth_1	sumoth_t	sumoth_r	sumoth_d	sumoth_e	sumriv_1	sumriv_t
sumother1	1.0000						
sumotherhpwt	0.9791	1.0000					
sumotherair	0.6948	0.7039	1.0000				
sumothermpd	0.9309	0.9341	0.7914	1.0000			
sumothersp_e	0.9902	0.9747	0.6335	0.8862	1.0000		
sumrival1	-0.3873	-0.3552	0.0832	-0.1527	-0.4667	1.0000	
sumrivalhpwt	-0.2744	-0.2163	0.1680	-0.0271	-0.3487	0.9532	1.0000
sumrivalair	-0.0227	0.0089	0.3275	0.2013	-0.1035	0.8830	0.9168
sumrivalmpd	-0.1400	-0.0923	0.2531	0.1132	-0.2131	0.9053	0.9456
sumrivalsp_e	-0.5178	-0.4797	-0.0277	-0.2790	-0.5909	0.9823	0.9356
	sumriv_r	sumriv_d	sumriv_e				
sumrivalair	1.0000						
sumrivalmpd	0.9281	1.0000					
sumrivalsp_e	0.8144	0.8576	1.0000				

We also report correlation coefficients among the instruments. We can confirm that the instruments are divided into two groups and that each group's instruments are highly correlated with each other.

6 Conclusion

In this article, we presented the CSA2SLS estimator and the corresponding command, `csa2sls`. The usage of `csa2sls` was illustrated with an empirical application. The Monte Carlo experiments show that 2SLS is biased when there are many instruments and that CSA2SLS outperforms 2SLS when instruments are correlated with each other. Because CSA2SLS is computationally intensive, an interesting future research question would be to develop a more efficient computation algorithm. An approach based on the stochastic gradient descent (see, for example, Lee et al. [2022]) can be a possible solution.

7 Acknowledgments

We would like to thank the editor and an anonymous reviewer for their valuable comments on this article and for their helpful feedback on the program code. Shin is grateful for partial support by the Social Sciences and Humanities Research Council of Canada (SSHRC-435-2021-0244).

8 Programs and supplemental material

To install the software files as they existed at the time of the publication of this article, type

```
. net sj 23-4
. net install st0732      (to install program files, if available)
. net get st0732          (to install ancillary files, if available)
```

9 References

- Berry, S., J. Levinsohn, and A. Pakes. 1995. Automobile prices in market equilibrium. *Econometrica* 63: 841–890. <https://doi.org/10.2307/2171802>.
- Donald, S. G., and W. K. Newey. 2001. Choosing the number of instruments. *Econometrica* 69: 1161–1191. <https://doi.org/10.1111/1468-0262.00238>.
- Kuersteiner, G., and R. Okui. 2010. Constructing optimal instruments by first-stage prediction averaging. *Econometrica* 78: 697–718. <https://doi.org/10.3982/ECTA7444>.
- Lee, S., Y. Liao, M. H. Seo, and Y. Shin. 2022. Fast and robust online inference with stochastic gradient descent via random scaling. In *Proceedings of the Thirty-Sixth International Joint Conference on Artificial Intelligence*, 7381–7389. Buenos Aires, Argentina: AAAI Press.
- Lee, S., and Y. Shin. 2021. Complete subset averaging with many instruments. *Econometrics Journal* 24: 290–314. <https://doi.org/10.1093/ectj/utaa033>.

About the authors

Seojeong Lee is an associate professor of economics at Seoul National University.

Siha Lee is an assistant professor of economics at McMaster University.

Julius Owusu is a doctoral candidate in economics at McMaster University.

Youngki Shin is a professor of economics at McMaster University.