Nearly Collinear Robust Procedures for 2SLS Estimation

Alwyn Young
London School of Economics
London, United Kingdom
a.young@lse.ac.uk

Abstract. Stata's two-stage least squares (2SLS) computation procedures are sensitive to near collinearity among regressors, allowing situations in which reported results depend upon factors as irrelevant as the order of the data and variables. This note illustrates this claim with the public use data of Oreopoulos, 2006, *American Economic Review* 96: 152-175, where by permuting the order of the variables the instrumented coefficient estimate can be made to vary between .012 and 30.0 in a single specification. Different methods for improving the accuracy of 2SLS estimates are reviewed and an ado file for collinearity robust 2SLS estimation in Stata provided.

1 Introduction

Users of Stata regularly rely on the programme's ability to weed out and drop perfectly collinear nuisance regressors. Problems arise, however, when regressors are not collinear enough to be flagged and dropped by Stata, but collinear enough to affect computational accuracy. When variables are nearly collinear floating point rounding errors in matrix operations are magnified and reported results become sensitive to factors as econometrically irrelevant as the order of the data and variables. Sensitivity to collinearity is greater when conditioning on nuisance variables substantively affects point estimates, i.e. precisely when otherwise irrelevant variables play an essential role in the regression by conditioning out potential bias. These issues are especially relevant for two stage least squares (2SLS) estimation, where the standard formula used by Stata's *ivregress* command incorrectly assumes that the estimated inverse of the matrix of instrument inner products times the matrix itself is exactly equal to the identity matrix.

This note proceeds as follows: Section 2 lays out the canonical formula for 2SLS estimation and how its implicit assumption of zero computational error in matrix inversion can render estimates sensitive to irrelevancies such as the order of the data and variables. Section 3 illustrates the problem using the public use data and instrumental variables regressions of Oreopoulos (2006). Estimated 2SLS coefficients in that paper using Stata's *ivreg* or *ivregress* commands are shown to be sensitive to econometrically irrelevant procedures, varying as much as from .012 to 30.0 in a single specification through a simple reordering of variables. Section 4 reviews various computational methods for 2SLS estimation and section 5 tests these, as well as the user written commands *ivreg2* (Baum et al 2010) and *xtivreg2* (Schaffer 2010), on Oreopoulos's data and a broad sample of published 2SLS regressions whose regressors are rotated to artificially increase collinearity. The user written commands are found to be much less sensitive to near collinearity than Stata's built-in routines, but still orders of magnitude more sensitive than can be achieved by partitioning the 2SLS regression, which avoids cascading matrix inverse errors and produces minimal sensitivity to econometrically irrelevant procedures. Section 6 introduces *pariv*, a Stata ado file that implements this collinearity robust 2SLS estimation method, checks the sensitivity of results to the order of the data and variables, and reports the maximum R^2 found in the regression of one instrument on the others. Section 7 concludes with suggestions for safer econometric and programming practice.

2 Typical 2SLS Estimation Methods

Instrumental variables estimates are usually implemented using the canonical textbook representation of two stage least squares. Following the notation of Stata's help files, let

$$\mathbf{y} = \mathbf{Y}\boldsymbol{\beta}_1 + \mathbf{X}_1\boldsymbol{\beta}_2 + \mathbf{u} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u} \quad \text{and} \quad \mathbf{Y} = \mathbf{X}_1\boldsymbol{\Pi}_1 + \mathbf{X}_2\boldsymbol{\Pi}_2 + \mathbf{V} = \mathbf{Z}\boldsymbol{\Pi} + \mathbf{V}$$

where **y** is the n x 1 vector of second stage outcomes, **Y** the n x p matrix of endogenous regressors, $\mathbf{X}_1$ the n x k_1 matrix of included instruments (exogenous regressors), $\mathbf{X}_2$ the n x k_2 matrix of excluded instruments, and **u** and **V** the n x 1 and n x p vector and matrix of second and first stage disturbances. The remaining (Greek) letters are vectors and matrices of parameters. Stata, as well as some of the toolboxes proffered online for users of Matlab, estimates the second stage coefficients using the formula[1]

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

Under normal circumstances, (1) is equivalent to running the OLS regression of **y** on the projection of **X** on **Z**, $\hat{\mathbf{X}} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}$. However, when **Z** is nearly collinear, $(\mathbf{Z}'\mathbf{Z})^{-1}$ as calculated using machine precision is not close to the true inverse of $\mathbf{Z}'\mathbf{Z}$, so the computed value of $(\mathbf{Z}'\mathbf{Z})^{-1}(\mathbf{Z}'\mathbf{Z})$ differs substantially from the identity matrix and (1) does not yield OLS coefficients of any sort. The error in the assumption that the computed value of $(\mathbf{Z}'\mathbf{Z})^{-1}(\mathbf{Z}'\mathbf{Z})$ equals the identity matrix will vary with the order of the data and variables, and even the processor, as these will affect the floating point error in the sums $\mathbf{Z}'\mathbf{Z}$ and the way in which these errors cumulatively affect the calculation of $(\mathbf{Z}'\mathbf{Z})^{-1}$. Estimated coefficients can then become substantively sensitive to what are otherwise econometrically irrelevant procedures.

3 An Illuminating Example: Oreopoulos (2006)

Oreopoulos (2006) estimates the Mincerian return to schooling using instrumental variables based on variation induced by compulsory schooling laws in the United Kingdom and (to a much lesser extent) the United States and Canada. Oreopoulos's IV specifications include quartic

[1]This is the formula given in Stata's on-line help entry for *ivregress*. Although the command's code is hidden from users (the ado file calls for the internal command *_regress*), I am to approximate the problematic results produced by *ivreg, ivregress* and *xtivreg* using a formula of this type, with demeaned variables when the regression contains a constant term (see sections 4 and 5 below).

polynomials in the age of the respondent at the time labour income is reported and/or quartic polynomials in the birth cohort as exogenous regressors (i.e. included instruments).[2] In all of the UK IV samples estimating a Mincerian return the R^2 of the projection of age on age raised to the 2^{nd} through 4^{th} power or cohort year on cohort year raised to the 2^{nd} through 4^{th} power is always in excess of .999998. The use of dummy variables for age, birth cohort, region, region interacted with birth cohort, and year further increases collinearity, with the maximum R^2 in the regression of one included instrument on the others lying above .99999989 in all Mincerian UK specifications. In sum, ancillary regressors, whose coefficients are not important enough to ever be reported, are highly collinear. Below I focus on the UK IV estimates of the Mincerian return, as these are by far the most sensitive.

Panel a of Table 1 lists all 15 IV estimated UK Mincerian returns and associated standard errors published in tables in the paper, as well as revised estimates posted on the AEA data page in 2008 by Oreopoulos in response to reported difficulties in reproducing his results. In panel b I replicate his results using the data and specifications given in his 2008 public use code, while randomly varying the order of the data 10000 times. As shown, the order of the data discernibly affects the estimated Mincerian return in most specifications, with max - min differences of .05 in a few cases. Panel c randomly permutes the order of the variables when entered in the regression. All estimated Mincerian returns are sensitive to this, with a max - min difference of up to 30. In all cases, the maximum and minimum coefficient estimates are associated with regressions in which Stata reports finite standard errors for the Mincerian return as well as for all other estimated coefficients (excluding variables that are dropped) and there is nothing in the results to alert the user to the fact that they are sensitive to what should otherwise be completely irrelevant procedures. With sample sizes in the thousands and dozens of dummies in some specifications, 10000 random permutations barely scratch the surface of the N! possible permutations of the order of the data or variables, understating the actual max-min difference. Percentiles, however, are more accurately estimated with random sampling, while providing a sense of the variation found in the typical permutation. As shown, the 5th to 95th percentile range of the estimated Mincerian return found in random permutations of variable order is in excess of .2 in three specifications and of .1 in five.[3]

Panel d of the table reports collinear robust estimates using the partitioned regression method described further below. I first use the original quartic specifications for age and birth cohort, showing results that differ only slightly from those reported in Oleopoulos's corrigendum, a

---

[2]Quartic age controls have become standard in this literature, appearing in, for example, Deveraux and Hart (2010) and Stephens and Yang (2014).

[3]To create counterfactual results that might have been reported, Table 1 replicates and permutes using the *ivreg* command and frequency weights of the paper's public use code. Results using the newer *ivregress* and frequency weights are virtually identical (see the on-line appendix). However, Oreopoulos (2006) should have used aweights, as the weights are the number of observations used to create the cell mean data (and aweights were used for similar US and Canadian regressions in the paper). Normally aweights and fweights produce identical results, but this is not the case with nearly collinear regressors. Tables 2 & 4 below change the code to aweights, to conform to the type of weights otherwise universally found in the large sample of IV papers examined therein. With aweights, *ivregress* produces substantially different and worse results than *ivreg*, as reported below.

Table 1. Instrumented Effect of a Year's Education on ln UK Labour Income (Oreopoulos 2006)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)
table/row/column	2/1/4	2/1/5	2/1/6	2/2/4	2/2/5	2/2/6	2/3/4	2/3/5	2/3/6	4/6/2	4/7/2	4/6/3	4/7/3	4/8/2	4/9/2
(a) reported results															
published	.147	.145	.149	.135	.187	.210	.174	.149	.148	.158	.195	.094	.066	.147	.150
2006	(.061)	(.063)	(.064)	(.071)	(.070)	(.135)	(.042)	(.044)	(.046)	(.049)	(.045)	(.057)	(.056)	(.061)	(.130)
revised	.112	.111	.125	.129	.180	.179	.041	.133	.135	.108	.053	-.056	-.032	.101	.110
2008	(.034)	(.033)	(.040)	(.076)	(.062)	(.096)	(.032)	(.027)	(.028)	(.033)	(.039)	(.047)	(.048)	(.042)	(.055)
(b) replicated coefficient range in 10000 random permutations of data order															
min	.091	.094	.100	.124	.177	.177	.036	.129	.127	.108	.054	-.056	-.032	.091	.100
5th percentile	.101	.106	.110	.127	.179	.178	.038	.133	.131	.108	.054	-.056	-.032	.098	.109
95th percentile	.122	.126	.129	.131	.182	.179	.043	.139	.137	.108	.054	-.055	-.032	.117	.129
max	.138	.144	.142	.133	.184	.179	.046	.144	.141	.108	.054	-.055	-.031	.141	.144
(c) replicated coefficient range in 10000 random permutations of variable order															
min	.091	-.018	-.007	.123	.082	.164	.021	.104	.055	.108	.053	-.056	-.035	.006	.012
5th percentile	.093	.078	.067	.125	.161	.176	.027	.122	.113	.108	.053	-.056	-.033	.061	.069
95th percentile	.176	.194	.298	.140	.196	.187	.057	.158	.172	.108	.054	-.055	-.031	.271	.287
max	.208	27.9	25.0	.141	5.80	2.81	.064	.264	13.3	.109	.056	-.054	-.027	8.80	30.0
(d) collinear robust estimates															
with	.111	.115	.119	.129	.181	.178	.040	.136	.134	.108	.054	-.055	-.032	.107	.118
quartics	(.033)	(.033)	(.039)	(.075)	(.060)	(.094)	(.031)	(.027)	(.027)	(.032)	(.038)	(.046)	(.047)	(.042)	(.056)
with	-.003	.026	.036	.199	.254	.264	-.003	.085	.092	.109	.055	-.059	-.038	.030	.026
cubics	(.032)	(.028)	(.044)	(.085)	(.076)	(.125)	(.028)	(.028)	(.029)	(.032)	(.038)	(.046)	(.047)	(.032)	(.038)
(e) 1 - maximum R^2 found in regressing one instrument on the others															
with quartics	$4.4e^{-8}$	$4.0e^{-8}$	$4.0e^{-8}$	$3.6e^{-8}$	$3.1e^{-8}$	$2.9e^{-8}$	$5.9e^{-8}$	$5.3e^{-8}$	$5.2e^{-8}$	9.5^{-8}	9.6^{-8}	$1.1e^{-7}$	$1.1e^{-7}$	$4.0e^{-8}$	$4.1e^{-8}$
with cubics	$1.3e^{-5}$	$1.1e^{-5}$	$1.1e^{-5}$	$9.0e^{-6}$	$6.7e^{-6}$	$6.4e^{-6}$	$1.3e^{-5}$	$1.1e^{-5}$	$1.1e^{-5}$	$1.4e^{-5}$	$1.4e^{-5}$	$1.6e^{-5}$	$1.7e^{-5}$	$1.1e^{-5}$	$1.1e^{-5}$

Notes: Replication and permutation using an Intel Xeon W-2175 CPU (results vary by brand of processor) and, following the public use code, using *ivreg* and frequency weights. Results using *ivregress* are all but identical (see footnote in text and on-line appendix). Table, row & column refer to location in original publication and revised to revised estimates and code posted by Oreopoulos in 2008. Standard error estimates in parentheses. Collinear robust estimates use partitioned IV regression as described below and are insensitive to data or variable order.

consequence of a fortuitous ordering of the polynomials (where sensitivity is greatest) in order of increasing power in the original specification. While Oreopoulos' highly collinear specifications illustrate the potential sensitivity of 2SLS results in Stata to econometrically irrelevant procedures, this sensitivity has no implications for the substantive interpretation of his results. Panel d also reports coefficient estimates using cubic specifications for the age and birth cohorts, which are much less collinear (panel e). When compared with the collinear robust results with the quartic, and the variation shown in panels b and c, these show that specifications that are sensitive to the ordering of the data or variables are those where conditioning on the near-collinear fourth order of the polynomials has a big effect on coefficient estimates. Specifications where conditioning on the quartic has little effect on the 2SLS estimates, such as those in columns (10) - (13), are relatively insensitive to data and variable order (panels b and c), despite having a degree of collinearity similar to that found in other specifications (panel e).

4 Nearly-Collinear-Robust 2SLS Procedures

This section considers alternative 2SLS computational procedures and methods for improving computational accuracy. As noted earlier, 2SLS estimates are often computed using the formula:

$$\text{Method A}: \ \hat{\boldsymbol{\beta}} = [\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}]^{-1}\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y}$$

When $\mathbf{Z}$ is nearly collinear the computed value of $\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}$ may not be close to $\hat{\mathbf{X}}'\hat{\mathbf{X}}$ and this approach does not actually calculate the OLS coefficients of a regression with predicted right-hand side values. An obvious solution is to force the computation of OLS coefficients, using the formula

$$\text{Method B}: \ \hat{\boldsymbol{\beta}} = (\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}'\mathbf{y}, \ \text{where } \ \hat{\mathbf{X}} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}$$

Unfortunately, when $\mathbf{Z}$ is nearly collinear the predicted values of included instruments $\hat{\mathbf{X}}_1 = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}_1$ differ from $\mathbf{X}_1$. A computationally more robust approach makes direct use of the fact that the predicted values $\hat{\mathbf{X}}_1$ should equal $\mathbf{X}_1$, computing the OLS estimates

$$\text{Method C}: \ \hat{\boldsymbol{\beta}} = (\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}'\mathbf{y}, \ \text{ where } \ \hat{\mathbf{X}} = [\hat{\mathbf{Y}}, \mathbf{X}_1] = [\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{Y}, \mathbf{X}_1]$$

This approach, however, reinserts estimates $\hat{\mathbf{Y}}$ based upon nearly collinear regressors $\mathbf{Z}$ alongside possibly nearly collinear regressors $\mathbf{X}_1$, repeating, with the addition of new variables, the estimation of a nearly collinear inverse, potentially magnifying computation errors. A better approach might be to make use of the partitioned regression given by

$$\text{Method D}: \ \hat{\boldsymbol{\beta}}_1 = (\hat{\tilde{\mathbf{Y}}}'\hat{\tilde{\mathbf{Y}}})^{-1}\hat{\tilde{\mathbf{Y}}}'\tilde{\mathbf{y}} \ \& \ \hat{\boldsymbol{\beta}}_2 = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'(\mathbf{y} - \mathbf{Y}\hat{\boldsymbol{\beta}}_1), \ \text{where } \hat{\tilde{\mathbf{Y}}} = \tilde{\mathbf{X}}_2(\tilde{\mathbf{X}}_2'\tilde{\mathbf{X}}_2)^{-1}\tilde{\mathbf{X}}_2'\tilde{\mathbf{Y}} \qquad (2)$$

and where ~ denotes residuals from the projection on $\mathbf{X}_1$, as in $\tilde{\mathbf{Y}} = \mathbf{Y} - \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{Y}$. Since

$$(\mathbf{Z}'\mathbf{Z})^{-1} = \begin{bmatrix} (\mathbf{X}_1'\mathbf{X}_1)^{-1} + (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2(\tilde{\mathbf{X}}_2'\tilde{\mathbf{X}}_2)^{-1}\mathbf{X}_2'\mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1} & -(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2(\tilde{\mathbf{X}}_2'\tilde{\mathbf{X}}_2)^{-1} \\ -(\tilde{\mathbf{X}}_2'\tilde{\mathbf{X}}_2)^{-1}\mathbf{X}_2'\mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1} & (\tilde{\mathbf{X}}_2'\tilde{\mathbf{X}}_2)^{-1} \end{bmatrix}$$

provides all of the inverses used in (2), implemention of Method D amounts to calculating the inverse of the nearly collinear matrix inverse $(\mathbf{Z}'\mathbf{Z})^{-1}$ once and only once.

One may also improve computational accuracy by not actually calculating matrix inverses. Many of the matrix operations in Methods A - D above involve calculating $\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$, where $\mathbf{x}$ and $\mathbf{b}$ are vectors and $\mathbf{A}$ a symmetric matrix. Rather than calculating the inverse, one can consider this as solving for $\mathbf{x}$ in the linear system $\mathbf{Ax} = \mathbf{b}$. Solutions of linear systems involve fewer calculations than matrix inversion and hence less opportunity for floating point errors to cumulate. When $\mathbf{A}$ is known to be symmetric positive-definite, use of the Cholesky decomposition $\mathbf{CC}' = \mathbf{A}$ further reduces the number of calculations needed (Press et al 2007). On the minus side, however, is the fact that solving $\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$ as the linear system $\mathbf{Ax} = \mathbf{b}$ for each instance of $\mathbf{b}$ implicitly allows the matrix inverse of $\mathbf{A}$ to vary across the calculations used in computing the 2SLS coefficients. As shown below, this becomes a consideration when the coefficients are already calculated with a high degree of accuracy using the matrix inverse approach.

Another way to improve computational accuracy is by improving the "conditioning" of matrices. In matrix algebra the condition number of a positive-definite matrix, the ratio of the largest to smallest eigenvalues, is a measure of the sensitivity of the solution for $\mathbf{x}$ in $\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$ to errors in the computation of $\mathbf{b}$ (Watkins 2002).[4] If we divide the matrix of instruments $\mathbf{Z}$ into $\mathbf{Z}_1$ and $\mathbf{Z}_2$,[5] then it is easily shown that the condition number of the matrix $\mathbf{Z}'\mathbf{Z}$ is always worse than that of $\mathbf{Z}_1'(\mathbf{I} - \mathbf{Z}_2(\mathbf{Z}_2'\mathbf{Z}_2)^{-1}\mathbf{Z}_2')\mathbf{Z}_1$, i.e. the matrix of residuals of $\mathbf{Z}_1$ projected on $\mathbf{Z}_2$. In addition, the dimensionality of the matrix of residuals is smaller, reducing the number of calculations. Consequently, provided $(\mathbf{Z}_2'\mathbf{Z}_2)^{-1}$ can be calculated exactly, and the coefficients associated with $\mathbf{Z}_2$ easily calculated given the coefficient estimates associated with $\mathbf{Z}_1$, partitioning the regression in this manner can improve accuracy. These conditions are satisfied when the regression contains a constant term or dummy variables, and I show below that demeaning the remaining variables greatly improves the accuracy of all of the methods described above.[6] Although the code for Stata's *ivregress* (earlier, *ivreg*) and *xtivreg* commands is not transparent,[7] as shown below their

[4] This condition number bounds errors for perturbations measured by vector norms. For component-wise perturbations, Skeel's (1979) condition number, equal to the largest of the row sums of $|A^{-1}||A|$, is more relevant.

[5] Not necessarily corresponding to the included and excluded instruments $\mathbf{X}_1$ and $\mathbf{X}_2$.

[6] Rescaling variables so that the diagonal of the matrix of inner products is the identity matrix is sometimes recommended (e.g. Gould 2018) and ensures that the condition number of the *K* x *K* matrix is less than or equal to *K* times the minimum condition number attainable by any form of rescaling (van der Sluis 1969). However, in practice it may worsen rather than improve the condition number, while not reducing the dimensionality of the remaining matrix calculations. As shown in the on-line appendix, when applied to the nearly-collinear data sets examined below, on average and in worse case outcomes on coefficient variability and bias it achieves much less than demeaning, and when applied in combination with demeaning (i.e. full standardization of the data set) does not improve on what is achieved by demeaning alone.

[7] As noted earlier, the ado files for these commands call for the internal command *_regress*.

sensitivity to near collinearity is on par with that found using method A with demeaned variables and matrix inverses rather than linear solutions.

5 Testing on Nearly Collinear Data Sets

For the purposes of testing the relative accuracy of the procedures described above, as well as that of Stata's built-in IV commands and supplemental user written routines, I draw upon a broad sample of Stata-based 2SLS regressions published in AEA journals examined in Young (2022). The sample covers 967 2SLS specifications in 29 papers[8] (including Oreopoulos 2006) and is restricted to regressions that have only one endogenous variable, as specifications with more than that were found to be exceedingly rare. 91 of these specifications have zero or one included instruments other than the constant term. As the rotation procedure I use below to increase collinearity requires more than one such instrument, these regressions are dropped, leaving 876 2SLS specifications, 39 in Oreopoulos and 837 in 28 other papers. As a summary measure of collinearity, I use the maximum partial R^2 (net of any absorbed fixed effects) found in the regression of one instrument in **Z** on the others (hereafter, R^{2Max}).[9]

For the sample described above, I randomly permute the order of the included instruments (other than absorbed fixed effects and the constant term) 50 times and calculate the coefficient of variation of the coefficient on the endogenous (instrumented) regressor using Stata's built-in estimation commands *ivregress* and *xtivreg*.[10] Figure 1 below graphs the logarithm of these

[8]As results vary with the processor used, all calculations in this paper were made using a single workstation (with an Intel Xeon W-2175 CPU). Because of this resource limitation, I dropped one paper with very large sample sizes and numbers of regressors from the Young 2022 sample. In addition, as *xtivreg* does not allow for weights, the weighted *xtivreg2* IV regressions of another paper were also dropped so as to allow for consistent comparison.

[9]$1 - R^{2Max}$ actually bounds the condition numbers. Let **A** denote the matrix of inner products of the instruments (demeaned by the constant or absorbed fixed effects), $\lambda_1 \geq ... \geq \lambda_k$ its ordered eigenvalues, and a_{ii} and b_{ii} the ith diagonal elements of **A** and $\mathbf{A}^{-1}$, respectively. As the partial R^2 of the regression of the ith instrument on the others is given by $1 - (a_{ii}b_{ii})^{-1}$, and by the Schur-Horn theorem $\lambda_1 \geq a_{ii}$ and $1/\lambda_k \geq b_{ii}$, we have $\lambda_1/\lambda_k \geq \max_i a_{ii}b_{ii} = (1 - R^{2Max})^{-1}$. As the maximum of the row sums of $|A^{-1}||A|$ is $\geq \max_i a_{ii}b_{ii}$, it also bounds the Skeel condition number. Since R^2s are not affected by rescaling of variables, $1- R^{2Max}$ bounds the lowest condition number attainable through rescaling **A** by post- and pre-multiplying by a diagonal matrix.
In the on-line appendix I show that $\log(1 - R^{2Max})$ explains about as much of the variation in the sensitivity of 2SLS results in Stata shown in the figures and tables below as the log regular or Skeel condition numbers of the matrix of demeaned instruments, even when this matrix is rescaled by its diagonal. As $\log(1 - R^{2Max})$ is highly collinear with the condition numbers, when entered together in the regression one or both measures are often statistically insignificant, but the condition numbers have the edge as they remain statistically significant about twice as often as $\log(1 - R^{2Max})$. Nevertheless, I base the discussion above on R^{2Max} as its scale and values are more meaningful to applied econometricians.

[10]Lest there be any confusion, it is worth emphasizing that these and later results are the moments of the computed estimates for a fixed data set. Kinal (1980) showed that with normal errors 2SLS only has moments of order less than or equal to the number of excluded instruments minus the number of endogenous variables. For the vast majority of published regressions examined here, which are exactly identified, this means that with normal errors the IV estimator has no moments at all. This result, however, refers to the moments of the estimator calculated across all realizations of the data generating process. In the figures and tables below I report the moments across permutations of variable order for a given (single) realization of the data generating process. As there are a finite number of such permutations, each yielding a finite point estimate, these moments always exist. 50 permutations

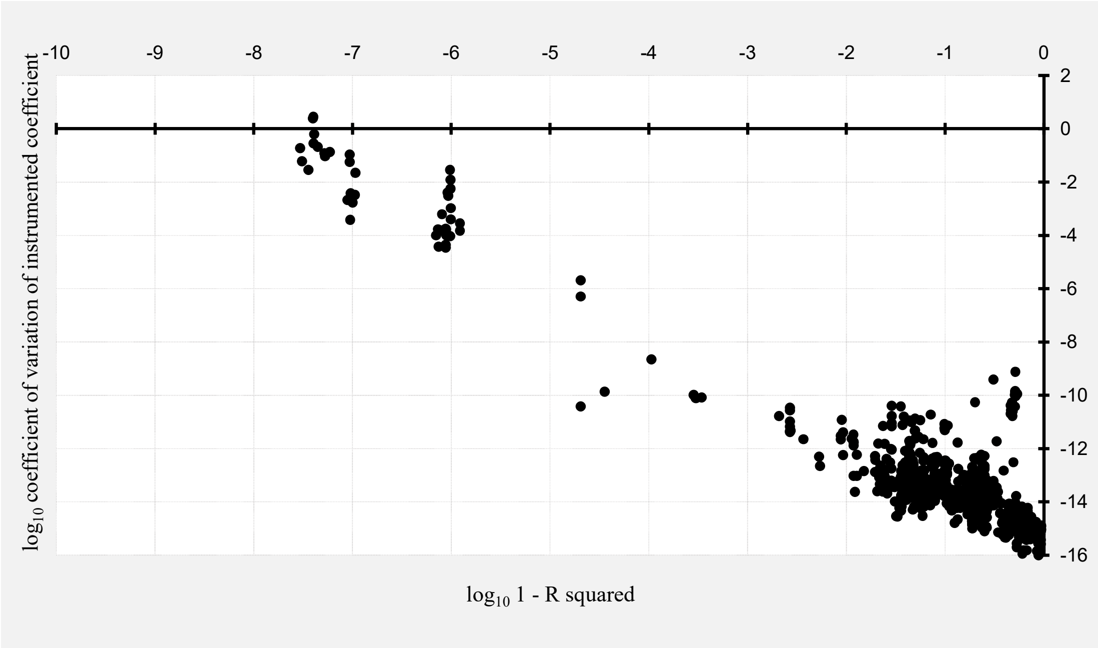

Figure 1: Coefficient of Variation of Instrumented Coefficients across 50 Permutations of Variable Order (876 IV regressions in 29 papers - 6 coefficients of variation are 0 and not shown in the figure)

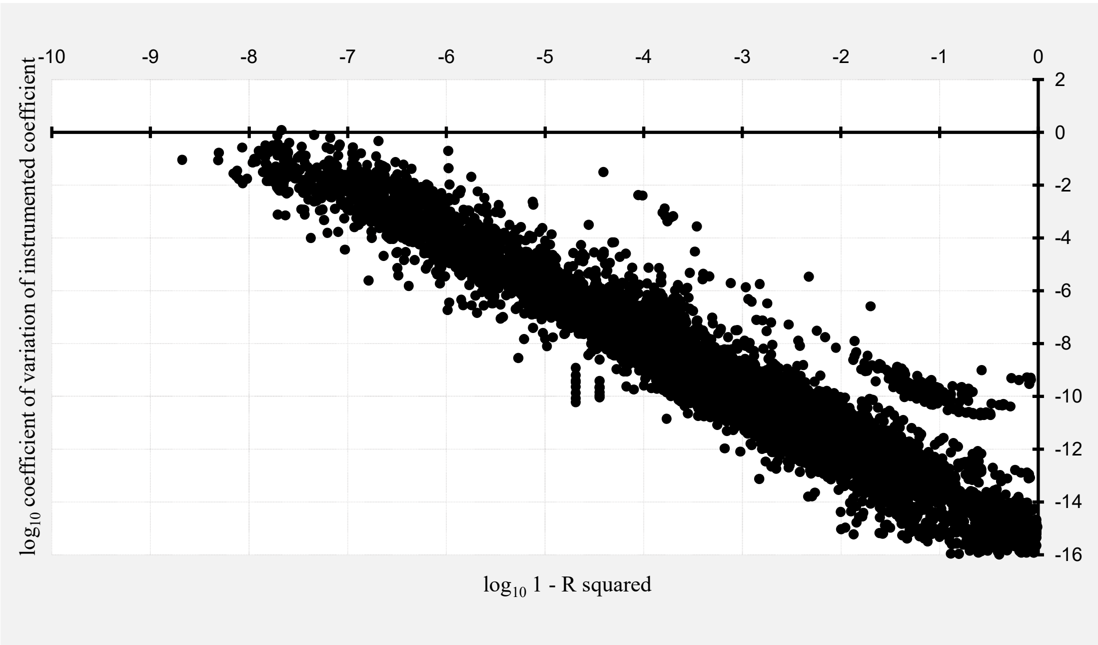

Figure 2: Coefficient of Variation of Instrumented Coefficients across 50 Permutations of Variable Order (8370 observations from 10 instrument rotations in 837 regressions in 28 papers - 32 coefficients of variation are 0 and not shown in the figure)

---

provide noisy estimates of these moments, but, when averaged across almost one thousand regressions in dozens of papers, are arguably enough to evaluate the relative accuracy of different computational methods.

against the logarithm in base 10 (to ease interpretation) of 1 - R^{2Max}. As shown, there is a strong relationship between the degree of collinearity and the coefficient of variation, but the sensitivity found in these papers (outside of Oreopoulos 2006 in the NE corner of the figure), while measurable, is not of substantive concern. For the purposes of testing alternative 2SLS computation procedures, I increase collinearity using a rotation procedure that theoretically, but not computationally, should be econometrically irrelevant.

For the N x k_1-1 matrix $\mathbf{X}_{1\sim c}$ of exogenous regressors other than the constant term and absorbed fixed effects, consider the rotation given by $\mathbf{X}^*_{1\sim c} = \mathbf{X}_{1\sim c}[\mathbf{U} + \mathrm{i} * \mathbf{I}(k_1 - 1)]$, where $\mathbf{U}$ is a k_1-1 x k_1-1 matrix of iid draws from the uniform distribution on (0,1), $\mathbf{I}(k_1\text{-}1)$ the k_1-1 dimensional identity matrix, and i is an integer scalar. $\mathbf{X}^*_{1\sim c}$ is often highly collinear, as each of the instruments is a linear function of the same k_1-1 variables, but $\mathbf{X}^*_{1\sim c}$ and $\mathbf{X}_{1\sim c}$ span exactly the same space. Consequently, as the excluded instruments are not included in $\mathbf{X}_{1\sim c}$, the rotation does not compromise the exclusion restriction and identification of the effect of the instrumented variable. After rotating $\mathbf{X}_{1\sim c}$ to $\mathbf{X}^*_{1\sim c}$ in each specification and calculating the new R^{2Max}, I then permute the order of the variables in $\mathbf{X}^*_{1\sim c}$ 50 times and calculate the coefficient of variation of the estimated coefficients across these permutations. With near-collinear regressors, Stata commands often drop regressors, turning nearly-collinear matrices into well-conditioned ones. The scalar i in $\mathbf{X}^*_{1\sim c} = \mathbf{X}_{1\sim c}[\mathbf{U} + \mathrm{i} * \mathbf{I}(k_1 - 1)]$ avoids this by reducing collinearity among the regressors. For each specification, I calculate $\mathbf{X}^*_{1\sim c}$ for each value of i = 1, 2, 3 ..., continuing up through the integers until I have 10 instances where in 50 permutations of variable order all regressors are retained by the Stata IV commands *ivreg* & *ivregress* (or, with fixed effects, *xtivreg*) and user written routine *ivreg2* (*xtivreg2*).[11] Thus, in each instance the variables are collinear, but not collinear enough to be flagged and dropped by existing routines.

Figure 2 graphs the $\log_{10}$ coefficient of variation using *ivregress* & *xtivreg* of the instrumented coefficient estimate across 50 permutations of instrument order against $\log_{10}(1\text{-}R^{2Max})$ for each of the 10 rotations of the included instruments in the 837 2SLS regressions of the 28 papers (excluding Oreopoulos 2006) in my sample. As shown, the rotations introduce a range of collinearity, with the highest degree of collinearity often generating a sensitivity similar to that found in Oreopoulos 2006, although there is considerable heterogeneity in the sensitivity of results in different papers to increasing collinearity. Regressions in the on-line appendix show that the coefficient of variation is increasing in the influence conditioning on the covariates has on the instrumented point estimate (as shown for Oreopoulos 2006 earlier above), but is not significantly related to the strength of the first stage or the number of observations or instruments.

Tables 2 and 3 below compare the accuracy of different 2SLS computational methods, with Table 2 focusing on original data and Table 3 on the nearly collinear samples created by the rotation procedure described above. Reported are results for Stata's *ivreg*, *ivregress* & *xtivreg*

[11] The fixed effects routines *xtivreg*/*xtivreg2* are used where authors did so in their code, which in all cases corresponds to situations where the large number of such effects makes inverting matrices which include them as dummy variables impractical.

commands, the user routines *ivreg2* & *xtivreg2*, and direct computation using the four methods described above in Stata's programming language Mata. Results labeled "demeaned" partition out the effect of the constant term, using the demeaned values of the remaining regressors, while those labeled "original" use calculations including the constant term in the matrix of regressors.[12] Results labeled "invert" invert matrices once and use them for all subsequent calculations, while those labeled "solve" compute each product of a matrix inverse with a vector as a separate Cholesky based solution of a linear system. Mean and maximum coefficients of variation are reported separately for the coefficients on the instrumented variable ($\hat{\boldsymbol{\beta}}_1$) and the included instruments ($\hat{\boldsymbol{\beta}}_2$).[13]

I begin by noting that Stata's native routines, *ivreg*, *ivregress* & *xtivreg*, are among the worst performing. True to Stata's documentation, these routines appear to use method A, producing results that are better than the rock bottom performance using method A with original data, but without achieving the improvements that come from demeaning, let alone the use of solvers rather than matrix inverses. *ivregress* produces results similar to the program it superseded, *ivreg*, but in Oreopoulos's (2006) data on average and in worst case situations (maxima) is worse.[14] The user written routines *ivreg2* and *xtivreg2* are orders of magnitude less sensitive than Stata's built in commands, producing results that are consistently very close to those found with method A and the use of linear solutions, but without (in the case of *ivreg2*) attaining the additional improvement found by demeaning.[15] Turning to the rest of the table, we see that calculations become systematically less sensitive to permutations of variable order as one moves from method A to B to C to D, and that demeaning and use of linear solutions confer large advantages, especially when using less accurate techniques such as method A. However, once method D is used, linear solvers actually appear to worsen accuracy since, as noted above, they implicitly allow the inverse of a given matrix to vary when solving different linear systems. Method D with demeaned variables is orders of magnitude less sensitive to collinearity than the user routines *ivreg2* and *xtivreg2*, although in the environments encountered in practical work (including Oreopoulos 2006) the latter are unlikely to display discernible sensitivity to variable order.

---

[12]In the case of three papers which use fixed effects estimation, the demeaned and original calculations are identical, as both make use of the remaining regressors net of absorbed fixed effects.

[13]When the data are demeaned, the coefficients on the constant term are recovered using the point estimates for the other effects and their coefficient of variation included in the reported figures. As the number of regressions and coefficients varies greatly by paper (see Young 2022), reported multi-paper means here and further below are calculated as the average of paper averages, so that each paper carries equal weight.

[14] Using aweights, as is done in this table. Using fweights, *ivreg* and *ivregress* produce virtually identical results, as noted earlier.

[15]*xtivreg2*, of course, demeans using the fixed effects. *ivreg2* and *xtivreg2*, based upon an examination of their ado files, make use of a mixture of matrix inverses and linear solutions.

Table 2. Mean and Maximum Coefficient of Variation of Coefficient Estimates across 50 Permutations of Variable Order (original data)

	mean				max			
	demeaned		original		demeaned		original	
	invert	solve	invert	solve	invert	solve	invert	solve
(a) Oreopoulos (2006) - $\hat{\boldsymbol{\beta}}_1$ - 39 coefficients on instrumented variables in 39 regressions								
ivreg	.17				2.3			
ivregress	.19				2.8			
ivreg2	2.2e-07				3.4e-06			
method A	7.8e-02	1.4e-08	1.6	2.3e-07	.70	2.3e-07	17	3.9e-06
method B	3.8e-09	5.1e-09	5.4e-08	5.4e-08	5.3e-08	8.9e-08	5.1e-07	5.1e-07
method C	8.3e-10	4.0e-09	2.8e-08	2.9e-08	8.1e-09	8.2e-08	1.6e-07	2.9e-07
method D	2.3e-13	2.3e-13	1.4e-11	5.5e-13	1.3e-12	1.2e-12	8.3e-11	2.4e-12
(b) Oreopoulos (2006) - $\hat{\boldsymbol{\beta}}_2$ - 2384 coefficients on included instruments in 39 regressions								
ivreg	8.7e-02				9.4			
ivregress	.12				38			
ivreg2	7.6e-07				5.5e-04			
method A	4.6e-02	3.2e-08	1.5	7.8e-07	12	3.5e-05	602	5.1e-04
method B	7.8e-09	1.3e-08	1.1e-07	1.1e-07	7.5e-06	1.4e-05	6.0e-05	6.1e-05
method C	1.0e-09	1.2e-08	4.2e-08	6.9e-08	6.9e-07	1.5e-05	2.2e-05	4.6e-05
method D	2.2e-10	7.5e-10	5.2e-09	1.1e-08	8.2e-08	2.3e-07	2.3e-06	4.1e-06
(c) 28 other papers - $\hat{\boldsymbol{\beta}}_1$ - 837 coefficients on instrumented variables in 837 regressions								
ivreg/xtivreg	6.6e-09				2.1e-06			
ivregress/xtivreg	6.6e-09				2.1e-06			
ivreg2/xtivreg2	1.4e-10				7.5e-09			
method A	7.2e-09	3.8e-12	6.8e-08	1.5e-10	2.3e-06	8.5e-10	4.9e-05	8.2e-09
method B	2.6e-12	2.0e-12	1.3e-10	6.3e-11	2.3e-10	3.9e-10	7.0e-09	3.1e-09
method C	2.2e-12	1.8e-12	1.3e-10	5.7e-11	1.8e-10	2.9e-10	6.9e-09	3.0e-09
method D	1.5e-15	1.4e-15	2.0e-15	2.0e-15	7.9e-14	7.0e-14	5.5e-13	6.3e-13
(d) 28 other papers - $\hat{\boldsymbol{\beta}}_2$ - 38263 coefficients on included instruments in 837 regressions								
ivreg/xtivreg	8.1e-09				1.3e-05			
ivregress/xtivreg	8.1e-09				1.3e-05			
ivreg2/xtivreg2	1.5e-09				9.0e-06			
method A	8.8e-09	2.2e-11	1.4e-07	1.4e-09	1.5e-05	1.1e-07	6.3e-04	8.6e-06
method B	2.3e-11	1.2e-11	1.2e-09	6.4e-10	1.1e-07	5.7e-08	7.0e-06	3.7e-06
method C	1.9e-11	1.2e-11	1.2e-09	6.0e-10	8.7e-08	6.0e-08	6.4e-06	3.6e-06
method D	6.7e-14	9.5e-14	3.5e-12	4.0e-12	5.8e-10	9.8e-10	6.9e-08	1.2e-07

Notes: Reported means for panels (c) and (d) are averages of paper averages, so that each paper carries equal weight. Coefficients on included instruments do not include *xtivreg/xtivreg2* absorbed fixed effects (in 3 papers). "demeaned" vs "original" - methods implemented using demeaned or original data, regressions with fixed effects always implemented using variables net of fixed effects; "invert" vs "solve" - matrix inverses using Mata's *invsym* function or products with matrix inverses solved using Mata's *cholsolve* function. 8 permutations for Oreopoulos (2006) where some of the exogenous regressors were dropped by Stata commands not included in calculations for any method. Instances in method A where any regressors dropped not included in calculations for that method alone. Other methods never drop a regressor.

Table 3. Mean and Maximum Coefficient of Variation of Coefficient Estimates across 50 Permutations of Variable Order for each of 10 Collinearity Increasing Rotations

	mean				max			
	demeaned		original		demeaned		original	
	invert	solve	invert	solve	invert	solve	invert	solve
(a) $\hat{\boldsymbol{\beta}}_1$ - 7930 coefficients in 10 data rotations of 793 regressions in 27 papers								
ivreg	4.5e-03				1.2			
ivregress	4.5e-03				1.2			
ivreg2	1.0e-08				5.3e-06			
method A	7.6e-02	7.8e-10	.10	1.1e-08	9.2	1.1e-07	926	4.6e-06
method B	1.6e-10	3.0e-10	2.6e-09	4.6e-09	8.2e-08	4.2e-08	1.5e-06	1.5e-06
method C	4.1e-11	2.7e-10	8.6e-10	4.0e-09	2.0e-08	4.6e-08	6.7e-07	1.2e-06
method D	1.8e-14	1.8e-14	2.5e-12	4.9e-14	7.1e-12	6.1e-12	2.0e-09	4.2e-11
(b) $\hat{\boldsymbol{\beta}}_2$ - 364540 coefficients in 10 data rotations of 793 regressions in 27 papers								
ivreg	5.8e-02				1502			
ivregress	5.8e-02				1905			
ivreg2	1.3e-07				4.2e-03			
method A	3.5e-02	3.0e-09	.32	1.2e-07	637	1.1e-04	4128	3.9e-03
method B	6.6e-09	1.5e-09	4.8e-08	6.0e-08	3.1e-03	5.8e-05	2.5e-03	2.7e-03
method C	2.0e-10	9.3e-10	8.8e-09	3.1e-08	2.1e-05	4.3e-05	3.5e-04	1.6e-03
method D	1.1e-10	4.9e-10	3.6e-09	2.5e-08	1.8e-05	3.6e-05	1.9e-04	1.4e-03
(c) $\hat{\boldsymbol{\beta}}_1$ - 440 coefficients in 10 data rotations of 44 regressions with fixed effects in 3 papers								
xtivreg	1.2e-03				.20			
xtivreg2	5.4e-09				2.5e-06			
method A	1.0e-03	5.7e-09			.19	2.8e-06		
method B	8.3e-10	1.1e-09	intrinsically		3.8e-07	5.0e-07	intrinsically	
method C	3.3e-10	8.5e-10	demeaned		1.1e-07	3.4e-07	demeaned	
method D	9.3e-15	9.5e-15			1.7e-12	2.0e-12		
(d) $\hat{\boldsymbol{\beta}}_2$ - 18090 coefficients in 10 data rotations of 44 regressions with fixed effects in 3 papers								
xtivreg	8.6e-03				4030			
xtivreg2	7.3e-09				5.9e-06			
method A	8.6e-03	7.6e-09			24	6.7e-06		
method B	1.5e-09	2.1e-09	intrinsically		1.2e-06	1.7e-06	intrinsically	
method C	4.9e-10	1.3e-09	demeaned		3.7e-07	7.9e-07	demeaned	
method D	1.4e-10	2.9e-10			3.5e-07	4.6e-07		

Notes: Reported means are averages of paper averages. For each of 837 regressions in 28 papers (excluding Oreopoulos 2006) collinearity randomly adjusted through variable rotations (in manner described in text) until 10 instances are found where all Stata commands don't drop a variable in 50 permutations of variable order. In method A instances where any regressors are dropped are not included in calculations for that method alone. Other methods never drop a regressor. Notation as in Table 2.

Lack of sensitivity to variable order is not equivalent to accuracy, as it is possible for a procedure to consistently provide incorrect estimates. To establish benchmark "true" values for the estimating equations in my sample, I calculate point estimates at 100 digit precision using the Advanpix Multiprecision Computing Toolbox for Matlab. When rounded to double precision, these estimates are identical across methods A through D implemented with and without demeaning or linear solutions for all 50 reorderings of Oreopoulos' regressors used in Table 2, i.e. display zero sensitivity to variable order. Using the 100 digit precision rounded to double precision values produced by the Advanpix Toolbox as the benchmark, Table 4 below reports the average relative absolute bias of the mean Stata coefficient estimates across 50 permutation of variable order. That is, with $\hat{\beta}_i(100)$ representing the 100 digit precision estimate rounded to double precision, and $\bar{\hat{\beta}}_i(method)$ the mean double precision point estimate across 50 random variable orders of methods A - D and the Stata routines in the tables, Table 4 reports the average and maximum value of $|\,[\,\bar{\hat{\beta}}_i(method) - \hat{\beta}_i(100)\,]/\hat{\beta}_i(100)\,|$ for the original data of Oreopoulos (2006) and the 10 collinearity increasing rotations of the data of the other 28 papers. The patterns follow the results of the previous tables, with Stata's built-in routines recording maximum relative bias as high as .99 on the instrumented coefficient and 913 on included instruments, *ivreg2* & *xtivreg2* lowering worst case outcomes to acceptable levels, and methods B through D successively lowering relative bias orders of magnitude further.

Table 4. Relative Bias across 50 Permutations of Variable Order

	mean				max			
	demeaned		original		demeaned		original	
	invert	solve	invert	solve	invert	solve	invert	solve
(a) Oreopoulos (2006) - $\hat{\boldsymbol{\beta}}_1$ - 39 coefficients in 39 regressions using original data								
ivreg	4.5e-02				.83			
ivregress	5.4e-02				.99			
ivreg2	7.6e-08				5.5e-07			
method A	2.3e-02	2.0e-09	.31	4.5e-08	.42	3.7e-08	2.8	8.9e-07
method B	9.0e-10	6.9e-10	1.1e-09	8.7e-09	1.1e-08	9.1e-09	1.1e-07	7.6e-08
method C	3.8e-09	1.0e-09	3.2e-08	2.8e-08	9.0e-08	6.7e-09	2.0e-07	1.5e-07
method D	6.8e-13	6.7e-13	4.6e-12	2.7e-13	4.6e-12	4.6e-12	3.6e-11	1.6e-12
(b) Oreopoulos (2006) - $\hat{\boldsymbol{\beta}}_2$ - 2384 coefficients in 39 regressions using original data								
ivreg	2.4e-02				14			
ivregress	5.3e-02				55			
ivreg2	7.9e-08				3.5e-05			
method A	1.5e-02	2.1e-09	.45	1.6e-07	11	3.9e-07	345	1.3e-04
method B	1.8e-09	7.0e-10	1.3e-08	9.8e-09	1.6e-06	1.4e-07	1.2e-05	4.2e-06
method C	1.3e-08	2.6e-09	4.7e-08	5.5e-08	1.8e-05	3.4e-06	6.2e-05	6.3e-05
method D	4.2e-10	3.2e-10	3.2e-09	4.3e-09	2.0e-07	1.5e-07	1.1e-06	1.3e-06
(c) 28 other papers - $\hat{\boldsymbol{\beta}}_1$ - 8370 coefficients in 10 data rotations each of 837 regressions								
ivreg/xtivreg	8.1e-04				.39			
ivregress/xtivreg	8.6e-04				.39			
ivreg2/xtivreg2	1.1e-09				6.5e-07			
method A	1.0e-03	2.2e-10	2.9e-03	1.7e-09	.96	6.0e-07	24	6.0e-07
method B	6.8e-11	6.1e-11	5.8e-10	6.0e-10	3.7e-08	3.9e-08	2.2e-07	2.6e-07
method C	1.6e-10	1.4e-10	1.0e-09	1.2e-09	2.6e-07	2.4e-07	8.1e-07	5.0e-07
method D	9.3e-15	9.5e-15	4.0e-13	1.1e-14	3.8e-12	3.8e-12	5.4e-10	2.8e-12
(d) 28 other papers - $\hat{\boldsymbol{\beta}}_2$ - 382630 coefficients in 10 data rotations each of 837 regressions								
ivreg/xtivreg	6.8e-03				913			
ivregress/xtivreg	6.9e-03				913			
ivreg2/xtivreg2	1.7e-08				1.1e-03			
method A	1.7e-02	5.3e-10	.12	1.3e-08	3990	8.0e-06	27770	7.1e-04
method B	2.1e-09	3.1e-10	6.8e-09	3.6e-09	1.4e-04	8.1e-06	1.9e-04	4.4e-04
method C	4.4e-10	4.5e-10	1.2e-08	1.2e-08	2.0e-05	9.4e-06	6.5e-04	4.7e-04
method D	1.9e-10	1.6e-10	5.4e-09	3.9e-09	1.7e-05	5.3e-06	5.6e-04	4.0e-04

Notes: Reported means in panels (c) and (d) are averages of paper averages. Otherwise, as in Tables 2 and 3.

6 *pariv*

For Stata users concerned about near collinearity in their IV regresion, *pariv* implements partitioned 2SLS (method D) using matrix inverses on demeaned data and if desired calculates the sensitivity of reported estimates to random permutations of data and variable order. The syntax and options are:

Syntax

pariv *depvar* (*endovars* = *excludedinst*) [*includedinst*] [*if*] [*in*] [*weight*] [,options]

Options

noconstant	no constant term
absorb(*varname*)	fixed effects for varname
small	finite sample adjustment of standard errors and degrees of freedom
robust	heteroskedasticity robust standard errors
cluster(*varname*)	clustered standard errors
reps(#)	number of permutations of data and variable order; default is 0
seed(#)	set random-number seed to #; default is 1

pariv fits the partitioned 2SLS regression of *depvar* on *endovars, includedinst* and (if specified) fixed effects for *varname* using *excludedinst* (as well as *includedinst* and any fixed effects) as instruments for *endovars*. To check that reported results are not substantively sensitive to econometrically irrelevant procedures, the user may call for reps(#) simultaneous permutations of data and variable order. *pariv* will then report the min to max range of the coefficient and standard error estimates of the partitioned regression across those permutations. *pariv* stores the following results in e():

e(Res)	Results table in matrix form.
e(ResB)	Coefficient estimates for each random permutation of data and variable order.
e(ResSE)	Standard error estimates for each random permutation of data and variable order.
e(R2max)	Maximum partial R2 found in regressing one instrument on the others.

The following code provides an illustrative example in which *ivregress*'s coefficient and standard error estimates depend heavily upon the order of the variables, but the collinear robust estimates produced by *pariv* do not (results for *ivregress* may vary with the processor used):

```
. drop _all
. set seed 836
. quietly set obs 16
. gen double age = _n + 19
. gen double age2 = age^2
. gen double age3 = age^3
. gen double age4 = age^4
. gen double u = invnormal(uniform())
. gen double e = invnormal(uniform())
. gen double z = invnormal(uniform())
```

```
. gen double t = 10*z + u
. gen double y = t + u + e

. ivregress 2sls y (t = z) age age2 age3 age4, robust


Instrumental variables 2SLS regression          Number of obs   =         16
                                                Wald chi2(5)    =    2790.55
                                                Prob > chi2     =     0.0000
                                                R-squared       =     0.9893
                                                Root MSE        =     1.0438

------------------------------------------------------------------------------
             |               Robust
           y | Coefficient  std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
           t |    .925502   .0553506    16.72   0.000     .8170169    1.033987
         age |  -264.6099   142.6179    -1.86   0.064    -544.1358      14.916
        age2 |   14.51373     7.8418     1.85   0.064    -.8559121    29.88338
        age3 |  -.3500912    .189302    -1.85   0.064    -.7211164     .020934
        age4 |   .0031352   .0016938     1.85   0.064    -.0001845    .0064549
       _cons |   1788.965   960.2967     1.86   0.062    -93.18207    3671.112
------------------------------------------------------------------------------
Instrumented: t
 Instruments: age age2 age3 age4 z


. ivregress 2sls y (t = z) age4 age age2 age3, robust


Instrumental variables 2SLS regression          Number of obs   =         16
                                                Wald chi2(5)    =    1111.02
                                                Prob > chi2     =     0.0000
                                                R-squared       =     0.9747
                                                Root MSE        =      1.609

------------------------------------------------------------------------------
             |               Robust
           y | Coefficient  std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
           t |   .7435555   .2981401     2.49   0.013     .1592116    1.327899
        age4 |   .0096079   .0103793     0.93   0.355    -.0107352     .029951
         age |  -795.1692   850.3027    -0.94   0.350    -2461.732    871.3935
        age2 |   43.94508   47.18243     0.93   0.352    -48.53079     136.421
        age3 |  -1.067046   1.149587    -0.93   0.353    -3.320196    1.186103
       _cons |   5332.821   5677.235     0.94   0.348    -5794.355       16460
------------------------------------------------------------------------------
Instrumented: t
 Instruments: age4 age age2 age3 z
```

```
. pariv y (t = z) age4 age age2 age3, robust reps(100)

Partitioned (collinear robust) 2SLS                Number of obs   =        16

                      Estimates               Statistical Significance
               coefficient  std. err.      |z|   P>|z|     [95% conf. interval]
-------------+---------------------------------------------------------------
           t |   .9395758  .04805926      19.55   0.000      .8453814    1.03377
        age4 |  .00263474  .00141279      1.86    0.062    -.00013428  .00540376
         age |  -223.5808   119.8144      1.87    0.062     -458.4128   11.25114
        age2 |   12.23788   6.572025      1.86    0.063     -.6430484   25.11881
        age3 |  -.2946544    .1582678     1.86    0.063     -.6048536  .01554485
       _cons |   1514.899   808.6364      1.87    0.061     -69.99973   3099.797
-----------------------------------------------------------------------------

     Range in 100 Permutations of Data and Variable Order
                     coefficients          standard errors
                    min        max          min        max
-------------+------------------------------------------
           t |   .9395758   .9395758    .04805926  .04805926
        age4 |  .00263474  .00263474    .00141279  .00141279
         age |  -223.5808  -223.5808     119.8144   119.8144
        age2 |   12.23788   12.23788     6.572025   6.572025
        age3 |  -.2946544  -.2946544     .1582678   .1582678
       _cons |   1514.899   1514.899     808.6364   808.6364
-------------------------------------------------------------
Instrumented: t
Excluded instruments: z
Included instruments: age4 age age2 age3 _cons
Heteroskedasticity robust standard errors

Maximum R2 found in the regression of any one instrument on the others:  .99999998
```

The minimum and maximum coefficient and standard error estimates are identical up to seven significant digits, and the user can be confident that the reported results are not substantively sensitive to econometrically irrelevant procedures.

7 Conclusion

The preceding results suggest that Stata users would do well to avoid Stata's native *ivregress* and *xtivreg* routines and make use of the computationally more accurate user written programs *ivreg2* and *xtivreg2*. At levels of near-collinearity that do not induce variable drops in either the original data of Oreopoulos 2006 or collinearity increasing rotations of the data of 28 other papers, these routines provide results which, within the range of typically reported significant figures, are accurate and totally insensitive to econometrically irrelevant procedures. For users who might nevertheless harbor concerns or curiosity, this paper provides *pariv* to check the sensitivity of the results and gauge (in the maximum R^2 of one variable projected on the others) the degree of near collinearity of the data. *pariv* is designed to be a confidence boosting check of computational accuracy and otherwise lacks the broad functionality found in *ivreg2* and *xtivreg2*.

Stata's computational methods are surprising, not least because Gould (2018), Stata's founding programmer, emphasizes the importance of computational accuracy and use of techniques such as demeaning and linear solutions which are clearly not consistently applied in Stata's IV code.

Additionally, the unnecessary and almost always incorrect assumption that a calculated matrix inverse times the matrix itself is exactly equal to the identity matrix (method A above), appears to be a defining feature of all Stata IV code, including user written routines. As Gould emphasizes, it is incumbent upon programmers to maximize computational accuracy, saving users the need to concern themselves with econometrically irrelevant issues. To this end, the tables above provide systematic evidence of the benefits of demeaning, linear solutions and partitioning of regressions in a broad practical sample.

8 References

Baum, C.F., M.E. Schaffer, S. Stillman. 2010. ivreg2: Stata module for extended instrumental variables/2SLS, GMM and AC/HAC, LIML and k-class regression. https://ideas.repec.org/c/boc/bocode/s425401.html.

Devereux, Paul, and Robert Hart. 2010. Forced to be Rich? Returns to Compulsory Schooling in Britain. *Economic Journal* 120: 1345-1364. https://doi.org/10.1111/j.1468-0297.2010.02365.x.

Gould, William. 2018. *The Mata Book: A Book for Serious Programmers and Those Who Want to Be*. College Station, TX: Stata Press.

Kinal, Terrence W. 1980. The existence of moments of k-class estimators. *Econometrica* 48: 241-249. https://doi.org/10.2307/1912027.

Oreopoulos, Philip. 2006. Estimating Average and Local Average Treatment Effects of Education When Compulsory Schooling Laws Really Matter. *American Economic Review* 96: 152-175. https://www.aeaweb.org/articles/pdf/doi/10.1257/000282806776157641.

Press, William H., Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery. 2007. *Numerical Recipes: The Art of Scientific Computing*. 3rd ed. Cambridge: Cambridge University Press.

Schaffer, M.E. 2010. xtivreg2: Stata module to perform extended IV/2SLS, GMM and AC/HAC, LIML and k-class regression for panel data models. https://ideas.repec.org/c/boc/bocode/s456501.html.

Skeel, Robert D. 1979. Scaling for Numerical Stability in Gaussian Elimination. *Journal of the Association for Computing Machinery* 26: 494-526. https://dl.acm.org/doi/pdf/10.1145/322139.322148.

van der Sluis, A. 1969. Condition Numbers and Equilibration of Matrices. *Numerische Mathematik* 14: 14-23. https://doi.org/10.1007/BF02165096.

Stephens, Melvin Jr., and Dou-Yan Yang. 2014. Compulsory Education and the Benefits of Schooling. *American Economic Review* 104: 1777-92. https://www.aeaweb.org/articles/pdf/doi/10.1257/aer.104.6.1777.

Watkins, David S. 2002. *Fundamentals of Matrix Computations*. 2nd ed. New York, NY: John Wiley and Sons.

Young, Alwyn. 2022. Consistency without Inference: Instrumental Variables in Practical Application. *European Economic Review* 147. https://doi.org/10.1016/ j.euroecorev.2022.104112.