Nearly Collinear Robust Procedures for 2SLS Estimation

Alwyn Young
London School of Economics, July 2022

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 (AER 2006), 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, reviews different methods for improving the accuracy of 2SLS estimates, and provides an ado file for collinearity robust 2SLS estimation in Stata.

I. 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 needlessly assumes that the estimated inverse of the matrix of instrument inner products times itself is exactly equal to the identity matrix.

This note proceeds as follows: Section II 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 III illustrates the problem using the public use data and instrumental variables regressions of Oreopoulos (2006). Stata's estimated 2SLS coefficients in that paper 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 IV reviews various computational methods for 2SLS estimation and Section V tests these on Oreopoulos's data and a broad sample of published 2SLS regressions whose regressors are rotated to artificially increase collinearity. Partitioning the 2SLS regression so as to avoid the repeated computation of matrix inverses is shown to be much more robust to near collinearity, producing virtually no sensitivity to econometrically irrelevant procedures. Section VI 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.

II. 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

$$(1)\quad \mathbf{y} = \mathbf{Y}\boldsymbol{\beta}_1 + \mathbf{X}_1\boldsymbol{\beta}_2 + \mathbf{u} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u} \quad \text{and} \quad \mathbf{Y} = \mathbf{X}_1\boldsymbol{\Pi}_1 + \mathbf{X}_2\boldsymbol{\Pi}_2 + \mathbf{V} = \mathbf{Z}\boldsymbol{\Pi} + \mathbf{V},$$

where $\mathbf{y}$ is the n x 1 vector of second stage outcomes, $\mathbf{Y}$ 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 $\mathbf{u}$ and $\mathbf{V}$ 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]

$$(2)\quad \hat{\boldsymbol{\beta}} = [\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}]^{-1}\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y}\,.$$

Under normal circumstances, (2) is equivalent to running the OLS regression of $\mathbf{y}$ on the projection of $\mathbf{X}$ on $\mathbf{Z}$, $\hat{\mathbf{X}} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}$. However, when $\mathbf{Z}$ 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 (2) 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

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

errors cumulatively affect the calculation of $\mathbf{(Z'Z)}^{-1}$. Estimated coefficients can then become substantively sensitive to what are otherwise econometrically irrelevant procedures.

III. 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' IV specifications include quartic 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 2nd through 4th power or cohort year on cohort year raised to the 2nd through 4th 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, ancilliary 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 I 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

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

Table I: 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
5^{th} percentile	.101	.106	.110	.127	.179	.178	.038	.133	.131	.108	.054	-.056	-.032	.098	.109
95^{th} 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	.163	.021	.104	.055	.108	.053	-.056	-.035	.006	.012
5^{th} percentile	.093	.078	.067	.125	.161	.176	.027	.122	.113	.108	.053	-.056	-.033	.061	.069
95^{th} percentile	.176	.194	.298	.140	.196	.184	.057	.158	.172	.108	.054	-.055	-.031	.271	.287
max	.208	27.9	25.0	.141	5.80	.701	.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)	(.034)	(.040)	(.076)	(.061)	(.095)	(.032)	(.027)	(.027)	(.033)	(.039)	(.047)	(.047)	(.043)	(.057)
with	-.003	.026	.036	.199	.254	.264	-.003	.085	.092	.109	.055	-.050	-.031	.030	.026
cubics	(.032)	(.029)	(.044)	(.087)	(.077)	(.127)	(.029)	(.028)	(.029)	(.033)	(.039)	(.046)	(.046)	(.033)	(.039)
(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: Standard error estimates in parentheses. Table, row & column refer to location in original publication and revised to revised estimates and code posted by Oreopoulos in 2008. Replication is done using Stata's *ivregress* command and an Intel i7 processor (results vary slightly by processor). Results using the discontinued *ivreg* command used in Oreopolous's code are virtually identical (on-line appendix). Collinear robust estimates use partitioned IV regression as described below and are insensitive to data or variable order.

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 minimum and maximum 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.

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 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 (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).

IV. Nearly-Collinear-Robust IV 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:

$$\text{(3) Method A: } \hat{\boldsymbol{\beta}} = [\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}]^{-1}\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y}\,.$$

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

$$\text{(4) Method B: } \hat{\boldsymbol{\beta}} = (\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}'\mathbf{y}, \text{ where } \hat{\mathbf{X}} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}\,.$$

Unfortunately, when $\mathbf{Z}$ is nearly collinear the predicted values $\hat{\mathbf{X}}_1 = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}_1$ 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

$$\text{(5) Method C: } \hat{\boldsymbol{\beta}} = (\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}'\mathbf{y}, \quad \text{where} \quad \hat{\mathbf{X}} = [\hat{\mathbf{Y}}, \mathbf{X}_1] = [\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{Y}, \mathbf{X}_1]\,.$$

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

$$\text{(6) Method D: } \hat{\boldsymbol{\beta}}_1 = (\hat{\tilde{\mathbf{Y}}}'\hat{\tilde{\mathbf{Y}}})^{-1}\hat{\tilde{\mathbf{Y}}}'\tilde{\mathbf{y}} \;\;\&\;\; \hat{\boldsymbol{\beta}}_2 = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'(\mathbf{y} - \mathbf{Y}\hat{\boldsymbol{\beta}}_1), \text{ where } \hat{\tilde{\mathbf{Y}}} = \tilde{\mathbf{X}}_2(\tilde{\mathbf{X}}_2'\tilde{\mathbf{X}}_2)^{-1}\tilde{\mathbf{X}}_2'\tilde{\mathbf{Y}}\,,$$

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

$$(7)\quad (\mathbf{Z}'\mathbf{Z})^{-1} = \begin{bmatrix} (\mathbf{X}_1'\mathbf{X}_1)^{-1} + (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2(\widetilde{\mathbf{X}}_2'\widetilde{\mathbf{X}}_2)^{-1}\mathbf{X}_2'\mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1} & -(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2(\widetilde{\mathbf{X}}_2'\widetilde{\mathbf{X}}_2)^{-1} \\ -(\widetilde{\mathbf{X}}_2'\widetilde{\mathbf{X}}_2)^{-1}\mathbf{X}_2'\mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1} & (\widetilde{\mathbf{X}}_2'\widetilde{\mathbf{X}}_2)^{-1} \end{bmatrix}$$

provides all of the inverses used in (6), 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). If we divide the matrix of instruments $\mathbf{Z}$ into $\mathbf{Z}_1$ and $\mathbf{Z}_2$,[3] 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 $\mathbf{Z}_2$. 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

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

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. Although the code for Stata's *ivregress* and (earlier) *ivreg* commands is not transparent,[4] I find I am able to closely reproduce the coefficient estimates for Oreopoulos' UK regressions produced by these commands, and their extraordinary variation with variable order, by implementing method A above with demeaned variables, using matrix inverses rather than linear solutions.

V. Testing on Nearly Collinear Data Sets

For the purposes of testing the relative accuracy of the procedures described above, I draw upon a broad sample of Stata-based 2SLS regressions published in AEA journals examined in Young (2022). The sample covers 1309 2SLS specifications in 30 papers (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 1218 2SLS specifications, 39 in Oreopoulos and 1279 in 29 other papers. As a summary measure of collinearity, I use the maximum partial R^2 (net of fixed effects or the constant term) found in the regression of one instrument in **Z** on the others (R^{2Max}).[5]

For the sample described above, I randomly permute the order of the included instruments

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

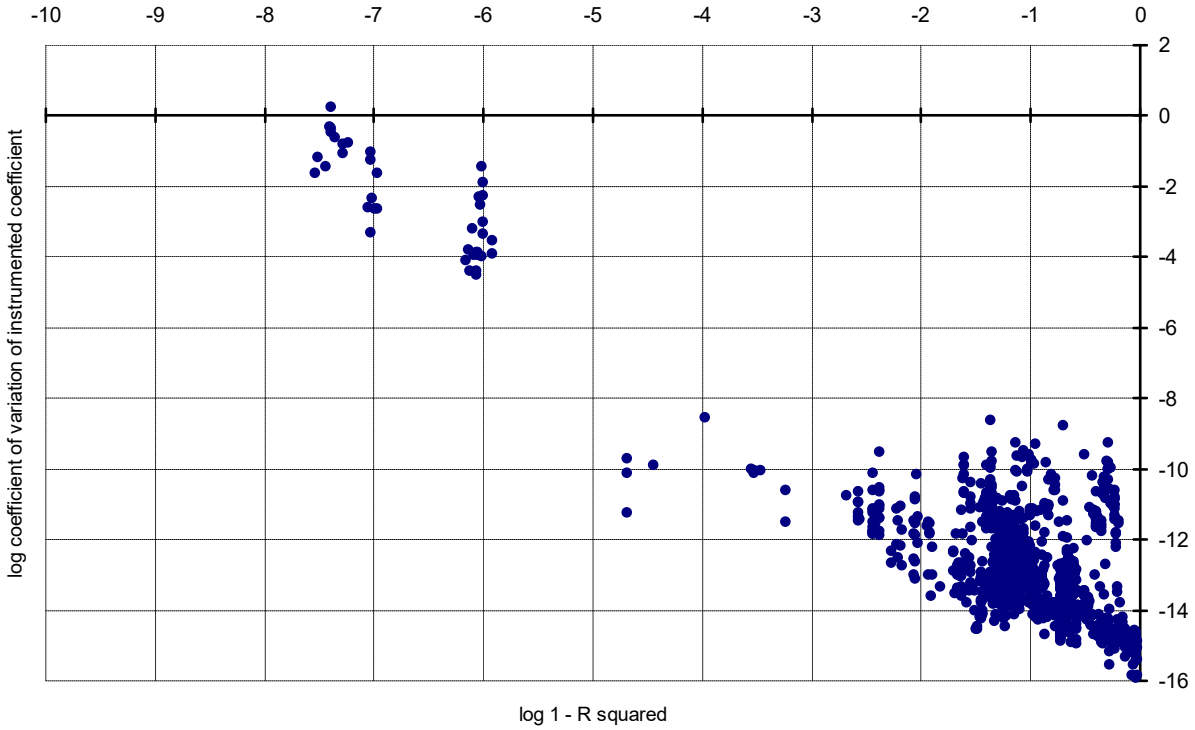
[5]In the on-line appendix I show that $\log(1 - R^{2Max})$ explains as much of the variation in the computational sensitivity of 2SLS results in Stata as the log condition number of the matrix of demeaned instruments, even when this matrix is rescaled by its diagonal to improve conditioning. This is not surprising, as the former bounds the latter. Let **A** denote the matrix of inner products of the instruments (demeaned by the constant or fixed effects), $\lambda_1 \geq ... \geq \lambda_k$ its ordered eigenvalues, and a_{ii} and b_{ii} the i^{th} diagonal elements of **A** and $\mathbf{A}^{-1}$, respectively. As the partial R^2 of the regression of the i^{th} 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 $\lambda_1/\lambda_k \geq \max_i a_{ii}b_{ii} = (1 - R^{2Max})^{-1}$. Since R^2s are not affected by rescaling of variables, this lower bound applies to the lowest condition number attainable by rescaling the matrix before inverting it.

(other than absorbed fixed effects and the constant term) 100 times and calculate the coefficient of variation of the coefficient on the endogenous (instrumented) regressor using each authors' chosen Stata estimation command (*ivregress*, *ivreg*, *ivreg2*, *xtivreg* or *xtivreg2*). Figure I below graphs the logarithm of this 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 $\mathbf{X}^*_{1\sim c} = \mathbf{X}_{1\sim c}(\mathbf{U} + i\mathbf{I}_{k_1-1})$, where $\mathbf{U}$ is a k_1-1 x k_1-1 matrix of iid draws from the uniform distribution on (0,1), $\mathbf{I}_{k_1-1}$ is the k_1-1 dimensional identity matrix, and i is an integer scalar. $\mathbf{X}^*_{1\sim c} = \mathbf{X}_{1\sim c}\mathbf{U}$ 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, the rotation should in principle have no effect on the estimated coefficient on the endogenous variable.[6] 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 $\mathbf{X}^*_{1\sim c}$ 100 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 $\mathbf{X}^*_{1\sim c} = \mathbf{X}_{1\sim c}(\mathbf{U} + i\mathbf{I}_{k_1-1})$ avoids this by reducing collinearity among the regressors. For each specification, I calculate one $\mathbf{X}^*_{1\sim c}$ for each value of i = 1, 2, 3 ..., continuing up through the integers until I have 10 instances where all regressors are retained by the authors' Stata IV command in all 100 permutations of the order of

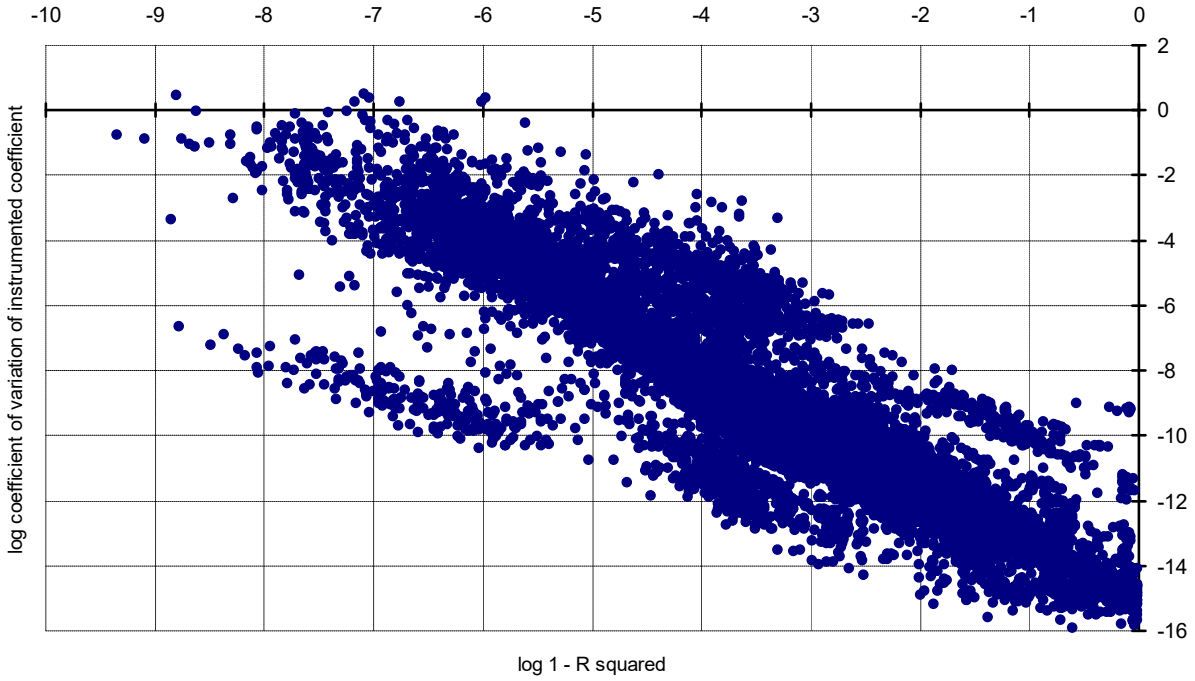
[6]Note that the excluded instruments are not included in $\mathbf{X}_{1\sim c}$.

Figure I: Sensitivity of Instrumented Coefficients to 100 Permutations of Instrument Order
(1218 2SLS regressions in 30 papers*)



*For 3 of the 1218 regressions the coefficient of variation is 0 and is not shown in the figure.

Figure II: Sensitivity of Instrumented Coefficients to 100 Permutations of Instrument Order
(11790 observations from 10 instrument rotations each of 1179 2SLS specifications in 29 papers*)



*For 8 rotations the coefficient of variation is 0 and is not shown in the figure.

the variables in $\mathbf{X}^{*}_{1\sim c}$. Figure II graphs the $\log_{10}$ coefficient of variation of the instrumented coefficient estimate across 100 permutations of instrument order against $\log_{10}(1\text{-}R^{2Max})$ for each of the 10 rotations of the included instruments in the 1179 2SLS regressions of the 29 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.

Table II below compares the accuracy of different 2SLS computational methods across the samples just described. Panel (i) focuses on Oreopoulos's 39 2SLS regressions, panel (ii) on the original 1179 specifications in the other 29 papers where collinearity is very much lower, and panel (iii) on the 11790 specifications arrived at by 10 collinearity increasing rotations of included instruments for each specification in the 29 papers. Reported are results for the Stata 2SLS commands called in authors' code and for direct computation using the four methods described above and 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.[7] 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

[7]In the case of four papers which use fixed effects estimation with *xtivreg*, the demeaned and original calculations both make use of the remaining regressors net of absorbed fixed effects, as the large number of fixed effects prevents calculations inverting the full matrix of regressors.

reported separately for the coefficients on the instrumented variable ($\hat{\boldsymbol{\beta}}_1$) and the included instruments ($\hat{\boldsymbol{\beta}}_2$).[8]

I begin by drawing attention to the equality between the results found using the *ivreg* command of Oreopoulos 2006 in the top row of panel (i), and those found using method A with demeaned variables and matrix inverses in the line below. As noted earlier above, this method reproduces quite closely the results produced by Stata's *ivreg* and *ivregress* commands as variables are reordered.[9] In panels (ii) and (iii) the reader will note that on average authors' Stata methods produce better results than method A with demeaned variables and matrix inverses. This is because author's methods include commands such as *ivreg2* and *xtivreg2* which, based upon an examination of their ado files, make use of a mixture of matrix inverses and linear solutions. These programmes achieve results that are better than demeaned method A with matrix inverses, but not as good as demeaned method A using linear solutions at all stages.

Turning to the systematic comparison of different approaches given for methods A through D, two patterns are apparent in the table. First, calculations become systematically less sensitive to permutations of variable order as one moves from method A to B to C to D. Second, techniques such as demeaning and linear solutions do not systematically improve computational accuracy once the baseline method is itself accurate enough that near collinearity ceases to be a problem. Using demeaned variables produces reductions in mean and maximum coefficients of variation relative to the use of original data of about 1 or 2 orders of magnitude in methods A

[8]Since the number of regressions and coefficients varies greatly by paper (see Young 2022), the means are calculated as the mean of paper means, so that each of the 29 papers carries an equal weight.

[9]With nearly collinear regressors, virtually every seeming irrelevant computational decision makes a substantial difference. One of the most consequential of these is how matrix cross-products are formed. I find that using the *matrix accum* command to form matrix cross-products allows me to closely approximate *ivregress* results using method A, while loading data into Mata and forming quad precision cross products there produces different and considerably worse results using method A. However, method B performs much better using the Mata cross product method, while calculations in methods C and D are considerably easier if this is done as well. Consequently, I implement methods B - D using quad precision cross products in Mata, and A using *matrix accum*.

Table II: Mean and Maximum Coefficient of Variation of Coefficient Estimates across 100 Permutations of Variable Order

	$\hat{\beta}_1$ - instrumented regressor								$\hat{\beta}_2$ - included instruments							
	mean				max				mean				max			
	demeaned		original		demeaned		original		demeaned		original		demeaned		original	
	invert	solve	invert	solve	invert	solve	invert	solve	invert	solve	invert	solve	invert	solve	invert	solve
authors' Stata command	(i) 39 2SLS specifications in Oreopoulos 2006 (original data)															
	1.0e-1				1.8				7.7e-2				4.4e+1			
method A	1.0e-1	1.2e-8	5.2	2.3e-7	1.8	1.9e-7	6.1e+1	3.9e-6	7.7e-2	3.3e-8	3.2	7.3e-7	4.4e+1	4.1e-5	5.2e+2	4.2e-4
method B	3.8e-9	5.1e-9	5.4e-8	5.4e-8	5.1e-8	8.3e-8	4.9e-7	4.9e-7	7.4e-9	1.3e-8	1.1e-7	1.1e-7	7.1e-6	1.5e-5	5.9e-5	5.5e-5
method C	8.4e-10	4.0e-9	2.7e-8	2.9e-8	8.1e-9	8.3e-8	1.5e-7	2.9e-7	1.0e-9	1.3e-8	4.2e-8	7.0e-8	7.1e-7	1.6e-5	2.5e-5	4.6e-5
method D	8.4e-10	2.3e-13	1.4e-11	5.5e-13	8.1e-9	1.2e-12	9.0e-11	2.6e-12	1.0e-9	7.7e-10	5.2e-9	1.1e-8	7.1e-7	2.5e-7	2.4e-6	4.7e-6
authors' Stata command	(ii) 1179 2SLS specifications in 29 other papers (original data)															
	2.1e-11				2.8e-9				2.2e-10				6.5e-7			
method A	5.7e-9	3.7e-12	7.3e-8	1.4e-10	1.8e-6	7.6e-10	5.2e-5	1.1e-8	7.0e-9	2.1e-11	1.4e-7	1.3e-9	1.2e-5	1.0e-7	6.3e-4	8.0e-6
method B	2.5e-12	2.2e-12	1.2e-10	7.9e-11	2.2e-10	3.9e-10	5.9e-9	9.7e-9	2.1e-11	1.3e-11	1.2e-9	7.6e-10	1.0e-7	6.0e-8	7.7e-6	4.8e-6
method C	2.3e-12	2.2e-12	1.2e-10	7.6e-11	2.1e-10	3.9e-10	7.0e-9	9.5e-9	1.9e-11	1.3e-11	1.1e-9	7.5e-10	9.7e-8	6.5e-8	6.7e-6	4.8e-6
method D	2.3e-12	2.4e-15	2.5e-15	2.5e-15	2.1e-10	4.2e-12	9.0e-13	9.8e-13	1.9e-11	1.0e-13	2.7e-12	3.7e-12	9.7e-8	9.4e-10	2.3e-8	3.0e-8
authors' Stata command	(iii) 1179 2SLS specifications in 29 papers (10 rotations of exogenous regressors for each specification)															
	3.4e-3				3.2				1.1e-1				1.9e+4			
method A	9.9e-3	1.8e-9	1.5e-1	1.7e-8	5.0e+1	2.8e-6	2.3e+2	1.6e-5	1.5e-1	2.2e-7	1.4	1.4e-6	1.9e+4	1.6e-1	4.1e+5	8.9e-1
method B	3.6e-10	5.8e-10	3.5e-9	6.5e-9	5.6e-7	6.4e-7	2.4e-6	6.8e-6	1.6e-7	1.1e-7	3.3e-7	3.8e-7	7.0e-2	8.5e-2	2.0e-1	2.0e-1
method C	1.2e-10	4.8e-10	1.1e-9	5.1e-9	1.9e-7	3.9e-7	7.0e-7	5.6e-6	1.7e-8	5.1e-8	6.7e-8	1.2e-7	1.2e-2	3.8e-2	4.1e-2	4.4e-2
method D	2.2e-14	2.2e-14	2.0e-11	5.9e-14	6.1e-11	5.5e-11	2.5e-8	1.4e-10	1.4e-8	3.2e-8	7.8e-8	1.4e-7	1.0e-2	2.4e-2	5.3e-2	7.1e-2

Notes: $\hat{\beta}_1$ (39 in Oreopoulos 2006, 1179 in 29 other papers); $\hat{\beta}_2$ (2384 in Oreopoulos 2006 and 96785 in 29 other papers - excluding absorbed fixed effects). "demeaned" vs "original": methods implemented using demeaned or original data, regressions with fixed effects (in 4 papers) always implemented using variables net of fixed effects; "invert" vs "solve" - symmetric matrix inverses calculated using Mata's *invsym* function or products with matrix inverses solved as the Cholesky solution of a system of linear equations using Mata's *cholsolve* function. Permutations in which any of the exogenous regressors are dropped by a method are not included in calculations for that method. This occurs rarely, and mostly when original data are used using method A. Collinearity in panel (iii) is adjusted so that authors' methods never drop a regressor in 100 permutations of variable order (see text above). Those same collinear regressors are then used for methods A-D.

through C, but no systematic benefits with method D. Using linear solutions produces reductions in coefficients of variation relative to matrix inverses in method A of multiple of orders of magnitude, but no systematic improvements once methods B through D are implemented.

As can be seen in Table III, using the partitioned regression approach provides large improvements over the 2SLS routines currently available in Stata. In panel (i) the mean and maximum coefficients of variation of .1 and 1.8 for instrumented coefficients and .077 and 44 for coefficients on included instruments found using Stata's *ivreg* (or *ivregress*) in Oreopoulos' 2SLS regressions are reduced to $1.4e^{-11}$ and $9.0e^{-11}$ and $5.2e^{-9}$ and $2.4e^{-6}$ using method D with original data and matrix inverses. In panel (iii), the mean and maximums of .0034 and 3.2 for instrumented coefficients and .11 and 19000 (!) for included instruments using Stata's diverse 2SLS commands in the collinearity increasing rotations for 29 papers are reduced to $2.0e^{-11}$ and $2.5e^{-8}$ and $7.8e^{-8}$ and $5.3e^{-2}$ using method D with original data and matrix inverses. Method D with matrix inverses is not only less sensitive to random permutations of variable order, it is also extremely accurate. For Oreopoulos's 15 highly collinear UK specifications I calculate coefficient estimates in Matlab with 100 digit precision using the Advanpix Multiprecision Computing Toolbox. Comparing these results, which when rounded to double precision are identical using all methods considered in Table III, to those arrived at using the partitioned regression with original data, matrix inverses and double precision computing in Stata, I find an average absolute difference in the estimated Mincerian return of $8.3e^{-12}$ and a maximum difference of $2.8e^{-11}$, while for the 385 coefficients on included instruments in those 2SLS regressions the average absolute difference is $5.5e^{-6}$ and the maximum difference $4.5e^{-4}$.

VI. *pariv*

pariv implements partitioned 2SLS using matrix inverses on the original data (unless fixed effects are called) 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.

Bibliography

Devereux, Paul, and Robert Hart (2010). "Forced to be Rich? Returns to Compulsory Schooling in Britain." *Economic Journal* 120 (549): 1345-1364.

Oreopoulos, Philip (2006). "Estimating Average and Local Average Treatment Effects of Education When Compulsory Schooling Laws Really Matter." *American Economic Review* 96 (1): 152-175.

Press, William H., Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery (2007). Numerical Recipes: The Art of Scientific Computing. Third edition. Cambridge: Cambridge University Press, 2007.

Stephens, Melvin Jr., and Dou-Yan Yang (2014). "Compulsory Education and the Benefits of Schooling." *American Economic Review* 104 (6): 1777-92.

Watkins, David S. (2002). Fundamentals of Matrix Computations. Second edition. New York: John Wiley and Sons, 2002.

Young, Alwyn (2022). "Consistency without Inference: Instrumental Variables in Practical Application." Forthcoming, *European Economic Review*.