Nearly Collinear Regressors and the Replicability and Robustness of 2SLS Results*

Alwyn Young
London School of Economics, December 2021

Abstract

In public use AEA code authors frequently include collinear nuisance regressors of no substantive interest, relying on software to select subsets of full rank. This approach renders estimates of parameters of principal interest sensitive to machine and programming tolerances, allowing situations in which published results depend upon factors as irrelevant as the order of the data and variables. I illustrate this claim for 2SLS with the public use data of Oreopoulos (AER 2006), develop guidelines on what degree of collinearity is problematic, and provide procedures and software for collinearity robust 2SLS estimation.

---

*I am grateful to Isaiah Andrews and Phil Oreopoulos for helpful comments.

I. Introduction

In public use AEA code authors frequently include collinear nuisance regressors of no substantive interest in their econometric specifications, relying on software to select a subset of full rank from the proffered list of variables.[1] Such procedures will sooner or later run afoul of machine and programme tolerances when variables are not collinear enough to be flagged and dropped by the programme 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 processor used or 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 needlessly assumes that the 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). Oleopoulos's coefficient estimates are shown to be substantively sensitive to econometrically irrelevant procedures, varying as much as from .012 to 30.0 in a single regression through a simple reordering of variables. This provides some explanation of the difficulties users of the public use data set have had in replicating his results. This note, however,

[1]For example, in my study of 31 AEA papers which use instrumental variables, Young (2021), I find that 11 papers construct a full set of dummies for a categorical variable and include them all in the regression along with the constant term, leaving it to Stata to sort out the collinearity.

has absolutely no implications for the assessment of Oreopoulos's results, as collinear robust estimates are close to those reported in the paper's corrigendum. Oreopoulos's paper merely highlights a risk AEA authors unknowingly face.[2] Section IV explores the question of what degree of collinearity is computationally problematic. I show that the problems encountered in Oreopoulos are measurably present, albeit not of a magnitude great enough to be of substantive concern, in a broad sample of 2SLS estimates published in AEA journals. Using another econometrically irrelevant procedure, i.e. a random rotation of the instruments that renders them more collinear, the instrumented coefficient estimates of these papers easily reproduce the sensitivity to variable order found in Oreopoulos 2006. A maximum R^2 in the regression of one instrument on the others of less than .99999 appears to be sufficient to ensure that in all but the most unusual cases reported estimates in Stata are not substantively sensitive to the order of the variables. Section V adopts a different approach to the problem, placing the burden on software rather than users. 2SLS estimation methods that do not needlessly assume zero error in the computation of matrix inverses and partition the regression so as to avoid the repeated computation of matrix inverses are shown to be much more robust to near collinearity, producing virtually no sensitivity to econometrically irrelevant procedures. An accompanying programme[3] for Stata users implements these collinearity robust 2SLS estimation methods, 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 VI concludes.

II. Typical 2SLS Estimation Methods

Instrumental variables estimates are usually implemented using the canonical textbook

[2]Oreopoulos' code is actually better than most, in that he makes use of the Stata *xi* command that alerts the programme to the use of dummies and automatically removes redundant categories for the full sample in memory (although the *xi* command produces collinear regressors when the regression uses a subset of the sample that does not include all values of the categorical variable).

[3]*ivpermute*, available on my website and (shortly) through Stata's command line *ssc install ivpermute*.

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[4]

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

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 $(\mathbf{Z}'\mathbf{Z})^{-1}(\mathbf{Z}'\mathbf{Z})$ differs substantially from the identity matrix and (2) does not yield OLS coefficients of any sort. The error in the assumption that $(\mathbf{Z}'\mathbf{Z})^{-1}(\mathbf{Z}'\mathbf{Z})$ equals the identity matrix will vary with the precision of the processor and the order of the data and variables, 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.

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

[4]This is the formula given in Stata's on-line help entry for *ivregress*. As the ado file for *ivregress* (as well as for the older *ivreg*) calls on *_regress*, an internal command whose code is not visible to users, it is not possible to confirm this, but I show further below that alternative formulae are much less sensitive than Stata's commands to near-collinearity.

(to a much lesser extent) the United States and Canada. The UK data provide a rare example where estimated local average treatment effects are close to average treatment effects and Oreopoulos finds returns that, in most specifications, are in the range of 10 to 18 percent. Using the same UK policy changes and data, Deveraux and Hart (2010) find average Mincerian returns of about 3 percent. Estimates using US compulsory schooling laws have similarly found both large (8 to 13 percent in Acemoglu and Angrist 2000) and small (zero or negative in Stephens and Yang 2014) returns.[5] This note has no implications for the substance of this literature, as computationally robust estimates below largely confirm Oreopoulos's reported results.

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).[6] 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, 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 of the table I attempt to replicate the results using the current public use data set and Stata code.

[5]I thank Phil Oreopoulos for bringing these varied results to my attention.

[6]Quartic age controls are not unusual 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
age/cohort quartic	no/yes	yes/yes	no/yes	no/yes	yes/yes	no/yes	no/yes	yes/yes	no/yes	yes/no	yes/no	yes/no	yes/no	yes/yes	yes/yes
dummies	no	no	yes	no	no	yes	no	no	yes	yes	yes	yes	yes	yes	yes
	(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 estimates using alternative processors														
Intel i7	.107	.124	.117	.127	.181	.178	.041	.133	.132	.108	.054	-.056	-.032	.107	.128
AMD 4000	.112	.111	.125	.129	.180	.178	.041	.134	.135	.108	.054	-.055	-.032	.101	.110
	(c) replicated coefficient range in 10000 random permutations of data order: Intel i7														
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
	(d) replicated coefficient range in 10000 random permutations of variable order: Intel i7														
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
	(e) collinear robust estimates														
revised 2008	.111	.115	.119	.129	.181	.178	.040	.136	.134	.108	.054	-.055	-.032	.107	.118
code	(.033)	(.034)	(.040)	(.076)	(.061)	(.095)	(.032)	(.027)	(.027)	(.033)	(.039)	(.047)	(.047)	(.043)	(.057)

Notes: Standard error estimates in parentheses. Table, row & column refer to location in original publication, dummies to whether dummy variables for age, cohort or data year are included, and revised to revised estimates and code posted by Oreopoulos in 2008. Collinear robust estimates use partitioned IV regression as described below and are insensitive to data or variable order. Replication is done using Stata's *ivreg* command, as in Oreopoulos' code. Results using the updated *ivregress* command are virtually identical (on-line appendix).

As shown, the Mincerian return varies by as much as .018 depending upon whether one uses a laptop with a typical Intel i7 processor or the AMD 4000 processor often found in gaming machines, despite the fact that the two processors use exactly the same samples, software and regressors. Panel c reports the range of coefficient estimates found using the Intel i7 in 10000 random permutations of the order of the data. 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 d 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 e at the bottom of the table reports collinear robust estimates, using the partitioned regression method described further below. These 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. Oreopoulos' highly collinear specifications illustrate the potential sensitivity of 2SLS results to

econometrically irrelevant procedures, but this sensitivity has no implications for the substantive interpretation of his results.

IV. Levels of Collinearity of Concern

The preceding raises the difficult question of what levels of collinearity should be of concern. This section uses a comprehensive sample of 2SLS regressions published in AEA journals to provide some perspective on this issue. As a metric of collinearity, I use one minus the maximum R^2 found in the regression of one instrument on the others. In computer science, the matrix "condition number", usually defined as the ratio of the largest to smallest eigenvalue, is often used as a metric of potential computation error. However, matrix inversion procedures, which are not visible to users of commercial software such as Stata, will change the reference matrix for which the condition number should be calculated.[7] There are also a variety of condition numbers, each providing appropriate error bounds for different operations (Higham 2002). I find that one minus the maximum R^2 explains as much, and often more, of the variation in the computational sensitivity of 2SLS results in Stata as various condition numbers (see the on-line appendix). Consequently, below I use the familiar R^2 as the metric of collinearity.[8]

As a practical sample of 2SLS regressions, I examine the 1359 2SLS regressions of 31 papers (including Oreopoulos 2006) published in AEA journals analyzed in Young (2021). The sample is comprehensive, using all papers identified by the keyword search "instrument" which provide Stata public use data files and code and use linear 2SLS procedures. Of the 1400 2SLS specifications identified in this manner, all but 41 had only one endogenous variable, and so the

[7]For example, a matrix may be rescaled before it is inverted to improve its condition or variables may be demeaned before the matrix of inner products is calculated. Both operations change the eigenvalues of the matrix of inner products.

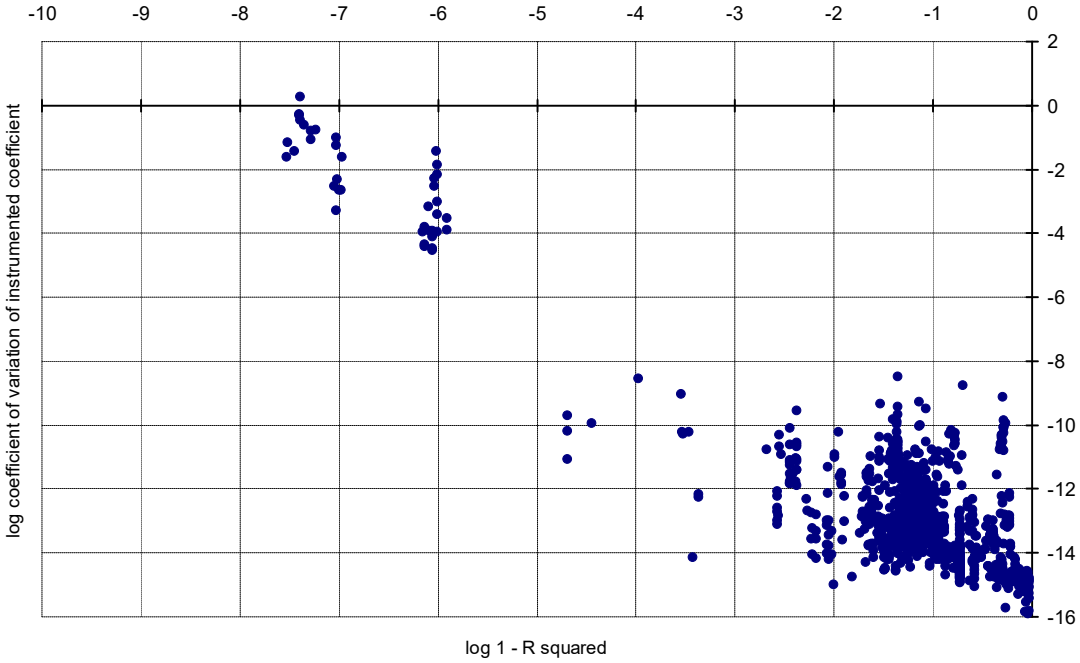
[8]One minus the maximum R^2 (R^{2Max}) actually bounds the inverse of the standard condition number of the matrix of inner products of the demeaned instruments. Let $\mathbf{A}$ denote this matrix, $\lambda_1 \geq ... \geq \lambda_k$ its ordered eigenvalues, and a_{ii} and b_{ii} the i^{th} diagonal elements of $\mathbf{A}$ and $\mathbf{A}^{-1}$, respectively. As the 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 $1 - R^{2Max} \geq \lambda_k/\lambda_1$.

analysis was restricted to specifications of that sort. Approximately 80 percent of the regressions contain only one excluded instrument. 137 regressions have zero or one included instruments other than the constant term. As the rotation procedure I use below to increase collinearity in the specification requires more than one such instrument, these regressions are dropped, leaving 1222 2SLS specifications found in 30 papers.

For the sample described above, I randomly permute the order of the instruments 100 times and calculate the coefficient of variation of the estimated coefficient on the endogenous (instrumented) regressor using the authors' designated Stata estimation command. Figure I graphs the logarithm of this (which is non-zero in all but three cases) against the logarithm of one minus the maximum R^2 (R^{2Max}) found in the regression of each of the instruments on the others. To ease interpretation, here and elsewhere below, the logarithm is in base 10. As shown in the figure, there appears to be a strong relationship between $1 - R^{2Max}$ and the coefficient of variation. It is, however, difficult to use the figure to draw guidance regarding levels of collinearity that may be problematic. All of the observations with a $\log_{10}(1- R^{2Max})$ less than -6 ($R^{2Max} > .999999$) belong to Oreopoulos 2006, most observations have a $\log_{10}(1- R^{2Max})$ greater than -3 ($R^{2Max} < .999$), and the few observations in between appear to lie below the regression line connecting the main sample to the Oreopoulos outliers, raising the question of whether the effects seen in the latter are unusual. Fortunately, an additional econometrically irrelevant procedure can resolve this issue by increasing the collinearity of the included instruments in the other published 2SLS specifications.

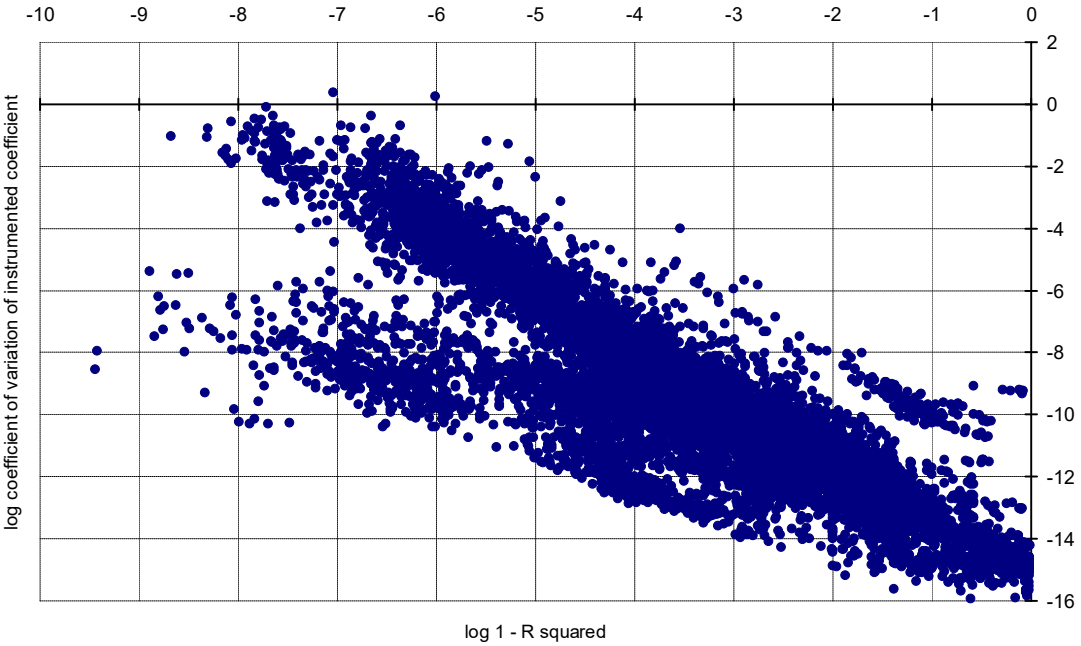
For the N x k_1-1 matrix $\mathbf{X}_{1\sim c}$ of exogenous regressors other than the constant term, consider the rotation given by $\mathbf{X}^*_{1\sim c} = \mathbf{X}_{1\sim c}(\mathbf{U} + i\mathbf{I}_{k_1-1})$, where $\mathbf{U}$ is a k_1-1 x k_1-1 matrix of iid draws from the uniform distribution on (0,1), $\mathbf{I}_{k_1-1}$ is the k_1-1 dimensional identity matrix, and i is an integer scalar. $\mathbf{X}^*_{1\sim c} = \mathbf{X}_{1\sim c}\mathbf{U}$ will often be highly collinear as each of the instruments is a

Figure I: Sensitivity of Instrumented Coefficients to 100 Permutations of Instrument Order
(1219 2SLS regressions in 30 papers*)



*For 3 of the 1222 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
(11812 observations from 10 instrument rotations each of 1182 2SLS specifications in 29 papers*)



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

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 have no effect on the estimated coefficient on the endogenous variable.[9] 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, revealing the sensitivity to collinearity of the given regression. 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 the 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 the variables in $\mathbf{X}^*_{1\sim c}$.

Figure II above graphs the $\log_{10}$ coefficient of variation of the instrumented coefficient estimate across 100 permutations of instrument order against $\log_{10}(1-R^{2Max})$ for each of the 10 rotations of the included instruments in the 1182 2SLS regressions of the 29 papers (excluding Oreopoulos 2006) in my sample.[10] This Figure confirms the downward sloping relationship seen in Figure I, but also indicates that there is considerable heteroskedasticity, with the variance of the outcome increasing with R^{2Max}. The outcomes seen in Figure I for Oreopoulos 2006 appear to be at the upper end of the distribution, but are also not unusual, as of the 1035 specifications in 24 papers with a $\log_{10}(1-R^{2Max})$ less than -6, 498 in 16 papers have a $\log_{10}$ coefficient of variation greater than -4. With results usually reported to 3 significant digits, a $\log_{10}$ coefficient of variation of -4 (i.e., .0001) is also a benchmark that ensures that reported results are not sensitive to random permutations of variable order. A maximum R^2 of .99999 appears to ensure that the

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

[10]In 8 cases with $R^{2Max} \leq .1$ the variation across 100 permutations is zero; these do not appear in the graph.

sensitivity of coefficient estimates rarely exceeds this bound in Stata.

Table II below regresses the $\log_{10}$ coefficients of variation shown in Figure II on the $\log_{10}(1\text{-}R^{2Max})$ of the random rotation and other characteristics of the regression specifications. Computation errors induced by collinearity among the included instruments are likely to pollute the 2SLS estimate of the instrumented coefficient more when conditioning on these is important for estimates. As a measure of the importance of conditioning on the included instruments, I use the proportional change in the estimated coefficient brought about by removing these, i.e. $\log_{10}|(\hat{\beta}_1 - \hat{\beta}_{1\sim\mathbf{X}_{1\sim c}})/\hat{\beta}_1|$, where $\hat{\beta}_1$ & $\hat{\beta}_{1\sim\mathbf{X}_{1\sim c}}$ are the estimated coefficients on the instrumented endogenous variable with and without the included instruments (other than the constant term) in the regression. Standard errors (in parentheses) are clustered at the 29 paper level with corrections for bias brought about by high leverage points, and p-values (reported below standard errors) adjusted for effective degrees of freedom based upon the volatility of standard error estimates created by these leverage points.[11]

As shown in Table II, $\log_{10}(1\text{-}R^{2Max})$ and $\log_{10}|(\hat{\beta}_1 - \hat{\beta}_{1\sim\mathbf{X}_{1\sim c}})/\hat{\beta}_1|$ by themselves explain about 3/4 of the variation in the log coefficient of variation across permutations of variable order, with the magnitude and precision of estimated effects increasing with the inclusion of paper or paper x specification fixed effects. In the last, the coefficient on $\log_{10}(1\text{-}R^{2Max})$ is estimated off of the variation induced by the rotation of regressors within each specification. The number of included instruments, number of observations, and 1st stage F of each specification are not robustly significant determinants of variation across permutations of variable order, with point estimates that are either statistically insignificant or change substantially with the inclusion of paper fixed effects.

[11]These reduce the statistical significance of reported results in the table and in general provide more accurate rejection probabilities than standard clustered/robust estimates and the hc corrections thereof (Young 2016).

Table II: Determinants of Log_{10} Coefficient of Variation
(11812 observations for 1182 2SLS specifications in 29 papers)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
$\log_{10}$ 1-R^{2Max}	-1.46 (.171) .000	-1.58 (.160) .000	-1.83 (.134) .000	-1.38 (.168) .000	-1.58 (.170) .000	-1.41 (.159) .000	-1.57 (.160) .000	-1.46 (.172) .000	-1.57 (.162) .000
$\log_{10}\left\|\frac{\hat{\beta}_1-\hat{\beta}_{1\sim\mathbf{x}_{1\sim c}}}{\hat{\beta}_1}\right\|$	.594 (.223) .025	.604 (.077) .000		.504 (.195) .030	.599 (.071) .000	.539 (.196) .022	.607 (.076) .000	.594 (.223) .026	.596 (.078) .000
$\log_{10}$ # of included instruments				.791 (.307) .031	-.254 (.525) .643				
$\log_{10}$ # of observations						.343 (.201) .126	.386 (.304) .244		
$\log_{10}$ 1st stage F								-.126 (.190) .527	-.372 (.292) .245
constant	-14.7 (.358) .000			-15.7 (.406) .000		-15.7 (.517) .000		-14.6 (.376) .000	
fixed effects		paper	paper x spec.		paper		paper		paper
R^2	.7522	.8987	.9704	.7772	.8988	.7682	.8990	.7530	.9008

Notes: Reported numbers = coefficient estimate, standard error estimate (in parentheses) clustered at paper level and adjusted for bias, & p-value with effective degrees of freedom corrections (last two as in Young 2016). R^{2Max} = maximum R^2 found in regression of the instruments on each other; $\hat{\beta}_1$ & $\hat{\beta}_{1\sim\mathbf{x}_{1\sim c}}$ = coefficient on instrumented regressor with included instruments (other than the constant term) and without these in the specification, respectively; spec. = 2SLS specification.

IV. Nearly-Collinear-Robust IV Procedures

This section considers alternative computational procedures and their effectiveness in reducing the sensitivity of coefficient estimates to near collinearity. Following the notation used in Section II above, Stata's method of calculating 2SLS estimates uses the formula:

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

When $\mathbf{Z}$ is nearly collinear $\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}$ 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

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

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

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

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

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

provides all of the inverses used in (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.

Table III below reports the average and maximum log coefficient of variation of 2SLS coefficients across 100 permutations of variable order. Panel (i) focuses on Oreopoulos's 40 2SLS regressions and panel (ii) on the 1182 specifications in the other 29 papers where collinearity is very much lower. The coefficients of variation of instrumented coefficients using Stata's commands for these were presented earlier in Figure I. Panel (iii) focuses on the 11820 specifications arrived at by 10 collinearity increasing rotations of included instruments for each of the specifications in the 29 papers. Coefficient of variations of the instrumented coefficients in these specifications using Stata commands were presented in Figure II. Reported in Table III 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 "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 ($\hat{\boldsymbol{\beta}}_2$). Logs are not used in summary statistics because in the case of less colinearity sensitive methods the coefficient of variation is frequently zero.

As shown in the table, method A with direct computation of matrix inverses produces results that are extremely sensitive to near collinearity, with an average coefficient of variation of of $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_2$ of 9.6 and 2.0, respectively, in Oreopoulos's 40 specifications and .067 and 1.2, respectively, in the 11820 colinearity-increasing rotations of the other papers' specifications.[12]

[12]These figures implicitly understate the variation, as the calculations in Table III exclude instances where any one of the original regressors is dropped (because of collinearity) after a permutation of variable order, which are frequent when using method A, but rare using the other methods (and never occur using method D).

Table III: Mean and Maximum Coefficient of Variation
of Individual Coefficient Estimates across100 Permutations of Variable Order

	$\hat{\boldsymbol{\beta}}_1$				$\hat{\boldsymbol{\beta}}_2$			
	mean		max		mean		max	
	invert	solve	invert	solve	invert	solve	invert	solve
(i) 40 2SLS specifications in Oreopoulos 2006 (original data)								
Stata command used by author	$1.0e^{-1}$		$1.8e^{0}$		$6.5e^{-2}$		$1.8e^{1}$	
A: $\hat{\boldsymbol{\beta}} = [\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}]^{-1}\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y}$	$9.6e^{0}$	$2.5e^{-7}$	$3.4e^{2}$	$4.8e^{-6}$	$2.0e^{0}$	$8.8e^{-7}$	$5.1e^{2}$	$5.5e^{-4}$
B: $\hat{\boldsymbol{\beta}} = (\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}'\mathbf{y},\ \hat{\mathbf{X}} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}$	$5.7e^{-6}$	$5.2e^{-6}$	$6.0e^{-5}$	$3.9e^{-5}$	$5.8e^{-6}$	$4.9e^{-6}$	$2.8e^{-3}$	$2.3e^{-3}$
C: $\hat{\boldsymbol{\beta}} = (\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}'\mathbf{y}, \hat{\mathbf{X}} = [\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{Y}, \mathbf{X}_1]$	$1.6e^{-7}$	$2.0e^{-7}$	$2.7e^{-6}$	$3.3e^{-6}$	$3.7e^{-7}$	$5.5e^{-7}$	$4.1e^{-4}$	$4.8e^{-4}$
D: $\hat{\boldsymbol{\beta}}_1 = (\hat{\tilde{\mathbf{Y}}}'\hat{\tilde{\mathbf{Y}}})^{-1}\hat{\tilde{\mathbf{Y}}}'\tilde{\mathbf{y}},\ \hat{\boldsymbol{\beta}}_2 = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'(\mathbf{y} - \mathbf{Y}\hat{\boldsymbol{\beta}}_1)$	$1.5e^{-11}$	$7.1e^{-13}$	$1.2e^{-10}$	$4.1e^{-12}$	$1.0e^{-8}$	$1.9e^{-8}$	$5.3e^{-6}$	$9.5e^{-6}$
(ii) 1182 2SLS specifications in 29 papers (original data)								
Stata commands used by authors	$1.4e^{-11}$		$3.3e^{-9}$		$4.5e^{-11}$		$7.9e^{-7}$	
A: $\hat{\boldsymbol{\beta}} = [\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}]^{-1}\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y}$	$1.7e^{-7}$	$7.6e^{-11}$	$1.1e^{-4}$	$1.2e^{-8}$	$9.7e^{-8}$	$3.9e^{-11}$	$1.3e^{-3}$	$2.3e^{-7}$
B: $\hat{\boldsymbol{\beta}} = (\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}'\mathbf{y},\ \hat{\mathbf{X}} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}$	$1.8e^{-9}$	$3.7e^{-10}$	$4.4e^{-7}$	$5.1e^{-8}$	$9.0e^{-10}$	$2.2e^{-10}$	$6.4e^{-6}$	$2.1e^{-6}$
C: $\hat{\boldsymbol{\beta}} = (\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}'\mathbf{y}, \hat{\mathbf{X}} = [\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{Y}, \mathbf{X}_1]$	$2.2e^{-10}$	$8.4e^{-11}$	$2.9e^{-8}$	$1.1e^{-8}$	$1.6e^{-10}$	$4.9e^{-11}$	$1.1e^{-6}$	$3.6e^{-7}$
D: $\hat{\boldsymbol{\beta}}_1 = (\hat{\tilde{\mathbf{Y}}}'\hat{\tilde{\mathbf{Y}}})^{-1}\hat{\tilde{\mathbf{Y}}}'\tilde{\mathbf{y}},\ \hat{\boldsymbol{\beta}}_2 = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'(\mathbf{y} - \mathbf{Y}\hat{\boldsymbol{\beta}}_1)$	$2.2e^{-14}$	$1.4e^{-14}$	$1.0e^{-11}$	$5.8e^{-12}$	$1.5e^{-11}$	$1.4e^{-11}$	$4.0e^{-7}$	$5.6e^{-7}$
(iii) 1182 2SLS specifications in 29 papers (with 10 rotations of exogenous regressors for each specification)								
Stata commands used by authors	$1.2e^{-3}$		$2.3e^{0}$		$7.5e^{-2}$		$1.9e^{4}$	
A: $\hat{\boldsymbol{\beta}} = [\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}]^{-1}\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y}$	$6.7e^{-2}$	$1.1e^{-8}$	$4.6e^{1}$	$1.8e^{-5}$	$1.2e^{0}$	$8.0e^{-7}$	$7.3e^{4}$	$6.0e^{-1}$
B: $\hat{\boldsymbol{\beta}} = (\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}'\mathbf{y},\ \hat{\mathbf{X}} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}$	$6.2e^{-8}$	$5.1e^{-8}$	$2.0e^{-4}$	$1.8e^{-4}$	$2.7e^{-6}$	$1.8e^{-6}$	$8.2e^{-1}$	$8.8e^{-1}$
C: $\hat{\boldsymbol{\beta}} = (\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}'\mathbf{y}, \hat{\mathbf{X}} = [\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{Y}, \mathbf{X}_1]$	$5.9e^{-9}$	$7.5e^{-9}$	$2.0e^{-5}$	$2.8e^{-5}$	$1.1e^{-7}$	$4.9e^{-7}$	$3.7e^{-2}$	$3.8e^{-1}$
D: $\hat{\boldsymbol{\beta}}_1 = (\hat{\tilde{\mathbf{Y}}}'\hat{\tilde{\mathbf{Y}}})^{-1}\hat{\tilde{\mathbf{Y}}}'\tilde{\mathbf{y}},\ \hat{\boldsymbol{\beta}}_2 = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'(\mathbf{y} - \mathbf{Y}\hat{\boldsymbol{\beta}}_1)$	$8.6e^{-12}$	$1.2e^{-13}$	$2.7e^{-8}$	$2.6e^{-10}$	$1.1e^{-7}$	$3.9e^{-7}$	$5.9e^{-2}$	$3.2e^{-1}$

Notes: $\hat{\boldsymbol{\beta}}_1$ = endogenous 2nd stage regressor (40 coefficients in Oreopoulos, 1182 in 29 papers); $\hat{\boldsymbol{\beta}}_2$ = exogenous 2nd stage regressor (2450 coefficients in Oreopoulos, 97363 in 29 papers - excluding absorbed fixed effects). "invert" - symmetric matrix inverses calculated using Mata's *invsym* function; "solve" - products of matrix inverses with vectors 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 are not included. No such drops occur for method D, but are frequent for method A.

However, the same method implemented using solutions of linear systems does much better, with average coefficients of variation of 10^{-7} or better in all panels, although a maximum value of .6 is attained in panel (iii) in the case of the coefficient on an included instrument. The sensitivity of Stata's own commands to variable reordering lies between the extremes of the matrix inversion and linear solution algorithms for this method, suggesting that Stata uses some combination of the two computational techniques.[13] Using matrix inversion, methods B through D successively reduce the sensitivity of coefficient estimates to random permutations of variable order, although improvements are less monotonic when using linear solutions.

As can be seen in Table III, the computational algorithm that is least sensitive to near-colinearity is method D, the partitioned regression. Using matrix inversion, average coefficients of variation for instrumented coefficients in panels (i) - (iii) are on the order of 10^{-11} or less, with maximums that never rise about 10^{-8}, while average and maximum coefficients of variation for the coefficients on included instruments never rise above the order of 10^{-7} and 10^{-2}, respectively. Linear solutions generally improve the averages for this method, but in panel (iii) also produce maximum (i.e. worst case) outcomes that are an order of magnitude worse than those found using 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 matrix inverses and double precision computing in Stata, I find an average absolute difference in the estimated Mincerian return of 8.3×10^{-12} and a maximum

[13] As noted earlier, since Stata's ado files for 2SLS call the internal command *_regress*, the exact 2SLS computational procedures used in Stata are not visible to users.

difference of 2.8×10^{-11}, while for the 385 ancilliary coefficient estimates in those 2SLS regressions the average absolute difference is 5.5×10^{-6} and the maximum difference 4.5×10^{-4}. Concern over such levels of precision seems ludicrous, but the accuracy given by programming in this fashion avoids the equally nonsensical sensitivity to permutations of data and variable order shown earlier in Table I and Figure II. The programme *ivpermute* implements partitioned 2SLS using matrix inverses, reports the maximum R^2 of the regression of the instruments on each other, and calculates the sensitivity of reported estimates to random permutations of data and variable order for users of Stata.

V. Conclusion

The alternative 2SLS algorithms reviewed above show the value of making software more robust to computationally challenging specifications. In particular, condensed textbook formulas, which assume that a matrix times its computed inverse equals the identity matrix, the projection of a variable on itself returns exactly the same variable, and iterative matrix inversion does not compound error, are best avoided. This note has focused on 2SLS, but such issues are very likely to arise in code for other econometric techniques as well. Consequently, users should probably, as a matter of course, avoid handing collinear or nearly collinear variables to software.

Bibliography

Acemoglu, Daron, and Joshua Angrist (2000). "How Large Are Human-Capital Externalities? Evidence from Compulsory Schooling Laws." NBER Macroeconomics Annual, Vol. 15, pp. 9-59.

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

Higham, Nicholas J. (2002). Accuracy and Stability of Numerical Algorithms. Second edition. Philadelphia: Society for Industrial and Applied Mathematics, 2002.

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.

Young, Alwyn (2016). "Improved, Nearly Exact, Statistical Inference with Robust and Clustered Covariance Matrices using Effective Degrees of Freedom Corrections." Manuscript, London School of Economics.

Young, Alwyn (2021). "Leverage, Heteroskedasticity and Instrumental Variables in Practical Application." Manuscript, London School of Economics.