

Replication Files for Weak Instruments in IV Regression: Theory and Practice

Isaiah Andrews, James Stock, and Liyang Sun

August 2, 2018

This readme documents the necessary steps to replicate all results in the paper and supplementary material of: “Weak Instruments in IV Regression: Theory and Practice” by Isaiah Andrews, James Stock, and Liyang Sun. Section A is an overview on the directory structure of the replication files. Section B describes features collected from articles and specifications in the AER sample.

A Directory Outline

Below is a description of the replication files directory. The order in which files are described below corresponds to the order in which files should be run to replicate results presented in the main text and supplementary material.

AER papers with IV estimates since 2014.xlsx collects information about articles and specifications in the AER sample. Section B details the features collected in this spreadsheet.

replication code To replicate the variance-covariance matrix for the reduced-form and first-stage estimates from our collected linear IV specifications, users should download replication files for each article from the AER website. The subdirectory name corresponds to the replication file name, which is also recorded in the spreadsheet. Within each subdirectory, the Stata script `/replications.do/` reads in the replication files and estimates the variance-covariance matrix maintaining whatever assumptions were used by the original authors (including the same controls, clustering at the same level, same degree-of-freedom adjustment and so on). We refer users to the replication files for documentation on the original specification.

replication results After running Stata script stored in `/replication code/`, the reduced-form and first-stage estimates should be stored in this directory. They are inputs to Matlab scripts that perform size simulations.

replication code homosked In this directory we provide Stata scripts `/replications.do/` that estimate the variance-covariance matrix assuming homoskedasticity for specifications with replication files. The structure of this directory is the same as `/replication code/`.

replication results homosked After running Stata script stored in `/replication code homosked/`, the reduced-form and first-stage estimates under homoskedasticity should be stored in this directory. While we do not present any result on this, one may be interested in size distortions when homoskedastic versions of first-stage F and robust tests are used. This can happen if we use the reported the Cragg-Donald statistic in the case of one endogenous variable, or use Stata packages that assume homoskedasticity to compute weak-IV robust tests such as `condivreg` when the data is in fact non-homoskedastic.

AER sample summary and AR test This R script reads in the spreadsheet and produces summary statistics on the AER sample. It also calculate AR confidence sets for just-identified specifications as described in the supplementary material.

batch.m This Matlab script reads in the spreadsheet and calls Matlab functions described below for each specification based on the specification id. Simulation results are written into `/simulation_eff_F/`.

olea_and_pflueger_eff_f_crit.txt We obtain critical values for the weak instrument test proposed by Montiel Olea and Pflueger (2013) from their paper and store the here for size simulations.

simulation.m This Matlab function calculates sizes based on simulations calibrated to AER sample as described in the supplementary material.

simulation_bayes.m This Matlab function calculates sizes using a Bayesian approach as described in the supplementary material.

simulation_overid.m This Matlab function calculates size of the overidentification test based on simulations calibrated to AER sample as described in the supplementary material.

collate_simulations_eff_F.m This Matlab script produces the figures presented in the main text and supplementary material. Figures are stored in `/figures/`.

B Collected Variables

Below we list features extracted from articles and specifications in the AER sample. These features are stored in the spreadsheet “AER papers with IV estimates since 2014.xlsx”. The order corresponds roughly to columns in the spreadsheet.

1. **specification** is the id assigned to a specification. We use this id as the key to store calibration and simulation results.
2. **Issue** is the AER issue in which the article is published in.

3. **Title** contains the article title.
4. **Author** contains the article author.
5. **Any linear IV estimates** indicates whether this specification estimates a linear IV model (1=yes;0=no).
6. **First stage specification** records which table in the article the first-stage is reported. It is left empty if the article does not report first-stage estimates.
7. **Number of instruments**
8. **Number of endogenous variables**
9. **First-stage coefficient(s) / standard error endo1 - endo4** collects the reported first-stage coefficient and standard error estimates. In specifications with multiple endogenous variables, there are multiple first-stage specifications, one for each endogenous variable. The columns correspond to endogenous variables in the order reported in the article.
10. **Any first-stage F reported** indicates whether any type of first-stage F-statistic is reported for the first-stage (1=yes;0=no).
11. **first-stage F statistic** collects the reported first-stage F-statistic. In specifications with multiple endogenous variables, there are multiple first-stage F-statistics, one for each endogenous variable. They are delimited by colon and listed in the order reported in the article.
12. **Any weak-IV-robust result** indicates whether any type of weak-IV-robust inference is used for this specification (1=yes;0=no). If so, then the test is recorded in the next column **What test**.
13. **Sample size** collects the number of observations used in the specification.
14. **reduced-form specification / coefficient / standard error** The **specification** column records which table in the article the reduced-form is reported. It is left empty if the article does not report reduced-form estimates or reduced-form estimates are reported in the same place as first-stage estimates. The rest columns collect the reported reduced-form coefficient and standard error estimates.
15. **IV specification / coefficient / standard error** The **specification** column records which table in the article the IV specification is reported. It is left empty if the article does not report IV estimates or IV estimates are reported in the same place as first-stage estimates. The rest columns collect the reported IV coefficient and standard error estimates.
16. **main specification** indicates whether this specification is the main specification, as defined in the supplementary material.
17. **Pre-test** describes what weak instrument test the specification entails, if any.

18. **public replication files / file name** indicates whether public replication files are available to replicate this specification. If so, then we record the replication file name when downloaded from the AER website. The file name is used to organize the `/replication code/` and `/replication code homosked/`.
19. **replicated** indicates whether we are able to replicate the variance-covariance matrix for the reduced-form and first-stage estimates by directly calculating them using the replication files as described in the supplementary material. Stata scripts for replication are stored in `/replication code/`.
20. **solved** indicates whether we are able to solve for the covariance estimate between reduced-form and first-stage based on 2SLS standard error estimate as described in the supplementary material.
21. **replicated first-stage / reduced-form variance / covariance** For specifications we solve for covariance estimates, these columns record the reported first-stage and reduced form estimates and variance to facilitate the calculation. For specifications we are able to replicate and just-identified, these columns record the replicated estimates to facilitate the calculation of AR confidence sets.
22. **replicated for comparison to homoskedasticity** indicates whether we are able to calculate the variance-covariance matrix assuming homoskedasticity. Stata scripts for estimation are stored in `/replication code homosked/`.
23. **notes on reported and actual first-stage F** indicate the type of F-statistics. The supplementary material describes how we categorize F-statistics.