Overidentification testing with weak instruments and heteroskedasticity

Stuart Lane 2025-09-25

Overview

This documentation provides an overview of the `oidrobust` for testing overidentifying restrictions in linear IV models. Sections 1 and 2 introduce the necessity of overidentification testing and introduce the model used in this empirical example (see Lane and Windmeijer (2025) for a more general exposition, and also a more detailed account of the empirical example in this document). Section 3 runs through the data and using the package with example code.

Packages

```
# New package for robust overidentification testing
devtools::install_github("stuart-lane/oidrobust", subdir="R")
library(oidrobust)

# Standard package used for IV estimation
library(ivmodel)

# Additional packages
library(dplyr)
library(readr)
```

1 Introduction

Overidentification tests allow users to assess the exogeneity of their instruments. These tests can be conducted whenever the number of instruments is greater than the number of endogenous regressors. As instrument exogeneity is a fundamental assumption, it is highly advised that users test their overidentifying restrictions whenever they have more instruments than endogenous regressors.

The `oidrobust` package offers improved overidentification testing in linear IV models. This package allows for the standard Sargan test (Sargan, 1958) and the Hansen J-test (Hansen, 1982), and allows for the use of overidentification tests based on the Limited Information Maximum Likelihood (LIML) estimator such as the Kleibergen-Paap test (Kleibergen & Paap, 2006), hereafter the KP-test. All test statistics considered are special cases of the score test (see Windmeijer (2021) and Lane and Windmeijer (2025) for theoretical properties and empirical evaluation of these test statistics).

In this document, we are going to replicate a small version of Table 6.1 of Lane & Windmeijer (2025) using the `oidrobust` package. Specifically, this paper argues that the KP-test is typically preferable for robust overidentification testing compared to the J-test.

2 Model

The model of interest is

$$\Delta c_{t+1} = \mu_c + \psi r_{t+1} + u_{t+1} \tag{1}$$

where Δc_{t+1} is the log of the consumption growth rate at time $t+1$, r_{t+1} is the log of the real interest rate at

time $t+1$, ψ is the parameter of interest, μ_c is a constant, and u_{t+1} is the unobserved error. The total number of observations is denoted T. Parameter ψ is the elasticity of intertemporal substitution. Here, r_{t+1} will be endogenous by construction (see Yogo (2004) or Lane and Windmeijer (2025) for details). However, u_{t+1} represents errors in expectations conditioned on the information set at time $t+1$, and therefore twice-lagged observable macroeconomic indicators should provide naturally valid instruments, in that $\mathbb{E}_t[Z_{t+1}u_{t+1}] = 0$, where Z_{t+1} is a matrix of instruments consisting of twice-lagged macroeconomic indicators and $\mathbb{E}_t[\,\cdot\,]$ denotes the expectations operator conditional on the information set at time $t+1$.

Since overidentification is easy to achieve in this model (simply select two macroeconomic indicators and twice-lag them), we should conduct an overidentification test to assess the validity of the instruments, we have good reason to believe *a priori* should be valid. The null and alternative hypotheses are:

$$H_0 : \mathbb{E}_t[Z_{t+1}u_{t+1}] = 0 \quad \text{v.s.} \quad H_1 : \mathbb{E}_t[Z_{t+1}u_{t+1}] \neq 0 \tag{2}$$

We can also re-normalise (1) as

$$r_{t+1} = \mu_r + \frac{1}{\psi}\Delta c_{t+1} + \eta_{t+1} \tag{3}$$

such that now r_{t+1} is the outcome of interest and Δc_{t+1} is now treated as the endogenous regressor. The moment restrictions now become

$$H_0 : \mathbb{E}_t[Z_{t+1}\eta_{t+1}] = 0 \quad \text{v.s.} \quad H_1 : \mathbb{E}_t[Z_{t+1}\eta_{t+1}] \neq 0 \tag{4}$$

where the null hypotheses in (2) and (4) are equivalent up to linear transformations, so validity of one moment restriction implies the validity of the other. However, we find empirically that weak instruments are a much bigger problem in the first specification in (3) than in (1).

2.1 Estimators and test statistics

The two estimators used in this empirical application are 2SLS and LIML, defined as:

$$\hat{\psi}_{2SLS} = \left((\Delta c)' P_z \Delta c\right)^{-1} \left((\Delta c)' P_z r\right)$$
$$\hat{\psi}_{L} = \left((\Delta c)' P_z \Delta c - \hat{\alpha}_L (\Delta c)' \Delta c\right)^{-1} \left((\Delta c)' P_z r - \hat{\alpha}_L (\Delta c)' r\right)$$

where observations have been stacked into vectors e.g. Δc is the stacked $T \times 1$ vector of observations Δc_{t+1}, $P_z = Z(Z'Z)^{-1}Z'$ for Z the $T \times k_z$ matrix of stacked instrument vectors Z'_{t+1}, and $\hat{\alpha}_L$ is the smallest root of the characteristic polynomial $|W'P_zW - \alpha W'W| = 0$ for $W = [\, r \;\; \Delta c \,]$. The score test for testing the hypotheses in (2), where $\hat{\psi}$ is either $\hat{\psi}_{2SLS}$ or $\hat{\psi}_L$, is given by

$$S(\hat{\psi}) = \hat{u}' M_{(\hat{\Delta c})} Z_2 \left(Z_2' M_{(\hat{\Delta c})} H_{\hat{u}} M_{(\hat{\Delta c})} Z_2 \right)^{-1} Z_2' M_{(\hat{\Delta c})} \hat{u}. \tag{5}$$

where $\hat{u} = r - (\Delta c)\hat{\psi}$, $(\hat{\Delta c}) = Z\hat{\pi}$ for the appropriate first-stage estimator $\hat{\pi}$, $M_{(\hat{\Delta c})} = I_T - (\hat{\Delta c})\left[(\hat{\Delta c})'(\hat{\Delta c})\right]^{-1}(\hat{\Delta c})'$, Z_2 are the overidentifying instruments, and $Z_2' M_{(\hat{\Delta c})} H_{\hat{u}} M_{(\hat{\Delta c})} Z_2 / T$ is some variance estimator. The statistic (5) nests numerous tests as special cases e.g. if $\hat{\psi} - = \hat{\psi}_{2SLS}$ and a homoskedastic variance estimator is used, then (5) is the standard Sargan test. If some robust variance estimator is used, then setting ψ to $\hat{\psi}_{2SLS}$ and $\hat{\psi}_L$ will give the Hansen J-test and KP-test respectively. Under standard (but technical) assumptions, when H_0 is true, then $S(\hat{\psi})$ will have a $\chi^2(k_z - 1)$ distribution in large samples, where k_z is the number of instruments.

2.2 Which test should I use?

Which test is better depends on numerous factors, and different models may lead to a different choice of the best test statistic for overidentification testing. Assuming strong instruments and homoskedasticity, both the standard Sargan test and a LIML-based variant perform similarly. Staiger and Stock (1997) recommend LIML-based testing due to greater robustness to weak instruments.

Lane and Windmeijer (2025) show that in general the KP-test performs better than the J-test used ubiquitously in empirical applications. Although both tests are numerically similar when instruments are strong, these tests can give very different answers when instruments may be weak and errors are non-homoskedastic. The KP-test is much less sensitive to large size distortions, and typically performs better than the J-test. There is also some evidence that the KP-test provides better power than the J-test when the instruments are strong. See Lane and Windmeijer (2025) for a detailed theoretical, numerical and empirical comparison of these test statistics.

3 Worked example

3.1 Data

The data used comes from Yogo (2004). This dataset consists of quarterly data on stock markets at the aggregate level, as well as macroeconomic variables from 11 countries: Australia (AUS), Canada (CAN), France (FRA), Germany (GER), Italy (ITA), Japan (JAP), Netherlands (NTH), Sweden (SWD), Switzerland (SWT), the United Kingdom (UK) and the United States of America (USA). The stock market data come from Morgan Stanley Capital International, and the consumption and interest rate data come from the International Financial Statistics of the International Monetary Fund.

For illustrative purposes, let's suppose we want to estimate ψ in (1) and the hypothesis in (2) for Australia (AUS). First, load the AUS data from the `AUSQ.txt` file.

```
# Load raw data from .csv file
df <- read_csv("../../Data/AUSQ.txt", col_names = TRUE, show_col_types = FALSE)
```

Let's get an overview of the dataframe by printing the head:

```
print(head(df))
```

```
## # A tibble: 6 × 12
##    DATE      r    dp    rf   inf     dc     rr    rrf    z1    z2    z3     z4
##   <dbl>  <dbl> <dbl> <dbl> <dbl>  <dbl>  <dbl>  <dbl> <dbl> <dbl> <dbl>  <dbl>
## 1 1970.  0.027 -3.87 0.013 0.006  0.007  0.021  0.007 -3.90 0.012 0.006  0.013
## 2 1970. -0.132 -3.63 0.014 0.017 -0.008 -0.149 -0.004 -3.85 0.012 0.018  0.001
## 3 1971.  0.014 -3.60 0.014 0.011  0.001  0.002  0.002 -3.87 0.013 0.006  0.007
## 4 1971. -0.054 -3.51 0.014 0.017  0.018 -0.071 -0.003 -3.63 0.014 0.017 -0.008
## 5 1971. -0.205 -3.29 0.013 0.017 -0.007 -0.221 -0.004 -3.60 0.014 0.011  0.001
## 6 1971.  0.167 -3.46 0.013 0.022 -0.001  0.145 -0.009 -3.51 0.014 0.017  0.018
```

The outcome of interest and the endogenous regressor are `dc` (the rate of consumption growth) and `rrf` (the real interest rate) respectively. The variables `z1`, `z2`, `z3` and `z4` are the instruments, which are simply twice-lagged versions of the log dividend-price ratio (`dp`), the nominal interest rate (`r`), the inflation rate (`inf`) and the rate of consumption growth respectively. See Yogo (2004) or Lane and Windmeijer (2025) for a more detailed description.

Define the variables from the dataframe `df`.

```
y <- df[['dc']]                      # outcome of interest (T x 1)
X <- df[['rrf']]                     # endogenous regressor (T x 1)
Z <- df[c('z1', 'z2', 'z3', 'z4')]   # instrument matrix (T x 4)
```

The 2SLS and LIML estimators can be computed using the standard `ivmodel` package:

```
iv_model <- ivmodel(Y = y, D = X, Z = Z)
coefficients <- coef(iv_model)

# Extract 2SLS estimator
results_2sls <- coefficients["TSLS", ]
psi_2sls <- results_2sls["Estimate"]

# Extract LIML estimator
results_liml <- coefficients["LIML", ]
psi_liml <- results_liml["Estimate"]

# Print the coefficents
cat(sprintf("2SLS estimate: %.2f, LIML estimate: %.2f\n", psi_2sls, psi_liml))
```

```
## 2SLS estimate: 0.05, LIML estimate: 0.03
```

We can also compute the 2SLS and LIML estimators, as well as the J- and KP-tests for validity of the overidentifying restrictions using the `score_test()` function in the `oidrobust` package. Here, we use the Newey-West heteroskedasticity and autocorrelation robust variance estimator (Newey & West, 1987) with 4 lags, although various types of variance estimator are available (see documentation). The tests can be implemented using user-defined matrices:

```
# Implementation of test statistics using matrices

# J-test computations
J_test <- score_test(y = y, X = X, Z = Z, W = NULL, method = "2sls", errors = "hac", lags =
4, no_constant = FALSE)

# KP-test computations
KP_test <- score_test(y = y, X = X, Z = Z, W = NULL, method = "liml", errors = "hac", lags =
4, no_constant = FALSE)
```

or alternatively can be implemented using a formula:

```
# Implementation of test statistics using formula

# J-test computations
J_test <- score_test(dc ~ rrf | z1 + z2 + z3 + z4, data = df,
                     method = "2sls", errors = "hac", lags = 4, no_constant = FALSE)

# KP-test computations
KP_test <- score_test(dc ~ rrf | z1 + z2 + z3 + z4, data = df,
                      method = "liml", errors = "hac", lags = 4, no_constant = FALSE)

#  Note - if we have a general dataset which includes:
#
#    - outcome of interest y
#    - K endogenous variables X1,..., XK
#    - L instruments Z1,..., ZL (L > K)
#    - P included exogenous variables W1,..., WP
#
#  then the general form of the formula required for the above is:
#
#      y ~ X1 + ... XK + W1 + ... + WP | Z1 + ... + ZL + W1 + ... + WP
```

From here, we then obtain the parameter estimates $\hat{\psi}_{2SLS}$ and $\hat{\psi}_L$ and the J- and KP-test statistics with

```
### 2SLS-based ========================

# Extract 2SLS estimate
psi_2sls <- J_test$coefficients

# Extract J-statistic value
J <- J_test$statistic["score"]

# Extract J p-value
J_p_val <- J_test$p.value

### LIML-based ========================

# Extract LIML estimate
psi_liml <- KP_test$coefficients

# Extract KP-statistic value
KP <- KP_test$statistic["score"]

# Extract KP p-value
KP_p_val <- KP_test$p.value
```

so collecting these estimates for Australia, and with the p-values in brackets, we find:

```
country_name <- "Australia"

cat(sprintf("%s| F: %.2f, cv: %.2f, 2SLS: %.2f, LIML: %.2f, J: %.2f (%.2f), KP: %.2f, (%.2f)
\n",
            country_name, 19.18, 18.40, psi_2sls, psi_liml, J, J_p_val, KP, KP_p_val)
    )
```

```
## Australia| F: 19.18, cv: 18.40, 2SLS: 0.05, LIML: 0.03, J: 8.78 (0.03), KP: 8.89, (0.03)
```

To see the effects of weak instruments, we also conduct an effective F-test and compute it's critical value (both values are pre-calculated here; see the replication files of Lane & Windmeijer (2025) for code to reproduce these). We see that the null of weak instruments cannot be rejected, and the statistic is in fact very low, indicating a potentially severe weak-instrument problem. It is clear that the KP-test does not reject the null hypothesis of valid instruments at the 5% level, whereas the J-test does. Given that the moment conditions in this model are *a priori* assumed valid, and that Lane and Windmeijer (2025) find that the J-test severely over-rejects the null under heteroskedastic weak instruments whereas the KP-test does not, we suggest that this is evidence of a false rejection of valid restrictions by the J-test rather than an incorrect failure to reject from the KP-test.

When we loop over both normalisations and all countries, we end up with the two tables: Table 1 estimates and tests the regression and hypotheses in (1) and (2) respectively, and Table 2 estimates and tests the regression and hypotheses in (3) and (4) respectively (note that the 95% critical value is 7.82 for the $\chi^2_{0.95}(3)$ distribution, where $\chi^2_{1-\alpha}(\lambda)$ is the $100(1-\alpha)$% critical value of the χ^2-distribution with λ degrees of freedom).

```
==================================================
                     TABLE 1
--------------------------------------------------
        F      cv     2SLS    LIML      J      KP
--------------------------------------------------
AUS   19.18   18.40   0.05    0.03    8.78    8.89
CAN   13.86   18.58  -0.30   -0.34    5.04    5.05
FRA   41.97   19.31  -0.08   -0.08    0.45    0.45
GER   13.37   18.32  -0.42   -0.44    2.59    2.54
ITA   21.44   18.92  -0.07   -0.07    1.07    1.06
JAP    5.44   21.29  -0.04   -0.05    4.73    4.73
NTH   12.18   18.52  -0.15   -0.14    3.69    3.69
SWD   21.19   18.76  -0.00   -0.00    2.59    2.59
SWT    7.90   18.03  -0.49   -0.50    2.25    2.27
UK     8.44   20.11   0.17    0.16    5.05    5.07
USA    8.14   18.21   0.06    0.03    7.14    7.58
--------------------------------------------------

==================================================
                     TABLE 2
--------------------------------------------------
        F      cv     2SLS    LIML      J      KP
--------------------------------------------------
AUS    2.47   19.49   0.50   30.03    9.49    8.89
CAN    2.98   18.07  -1.04   -2.98    6.96    5.05
FRA    0.22   19.67  -3.12  -12.38    2.08    0.45
GER    1.13   18.59  -1.05   -2.29    3.16    2.54
ITA    0.49   18.89  -3.34  -14.81    3.99    1.06
JAP    1.98   17.89  -0.18  -21.56    8.42    4.73
NTH    1.67   19.15  -0.53   -6.94    9.91    3.69
SWD    0.87   17.28  -0.10 -399.86   13.28    2.59
SWT    1.58   19.85  -1.56   -2.00    2.92    2.27
UK     2.68   17.62   1.06    6.21    8.17    5.07
USA    2.65   17.61   0.68   34.11    9.84    7.58
--------------------------------------------------
```

If the user wants to see the critical value in the general case, then they can run

```
# Obtain degrees of freedom
dof <- J_test$parameter["dof"]

# or equivalently
dof <- KP_test$parameter["dof"]

# and then compute the 1 - alpha (e.g. alpha = 0.05) critical value as
alpha = 0.05

critical_value <- qchisq(1 - alpha, df = dof)

sprintf("%.0f%% critical value of chi-squared distribution with %.0f degrees of freedom : %.3
f",
        (1 - alpha) *100, dof, critical_value)
```

```
## [1] "95% critical value of chi-squared distribution with 3 degrees of freedom : 7.815"
```

3.2 Interpretation of results

In Table 1, we see that the J and KP-tests agree on whether to reject the null in each of the 11 specifications considered. Both tests reject the null hypothesis at the conventional 5% level for Australia, but fail to find sufficient evidence to reject the null at this level for all other specifications. Although a number of these specifications fail to reject the null of weak instruments, the effective F-statistics are broadly large enough to likely conclude that the instruments have some explanatory power, although there may be some small weak-instrument bias. Overall, these results are consistent with the *a priori* belief that the instruments should be valid in this model, and the one rejection in 11 specifications is consistent with a 5% false rejection frequency.

In Table 2 however, we see that instruments are much weaker. The J-test rejects in 6 of the 11 specifications, but the KP-test continues to reject in just one specification (on inspection, one notes that for each country, the LIML estimates in Table 1 and Table 2 are exact inverses on each other and the KP-statistic is equivalent, but this is not the case for 2SLS and J). This is important for a couple of reasons. Firstly, we have good *a priori* reasons to believe that the instruments are valid. Lane and Windmeijer (2025) demonstrate that the J-test can suffer from a severe over-rejection problem with heteroskedastic weak instruments, something that is not found with the KP-test. Given that either the instruments tested in Tables 1 and 2 are equivalent up to linear transformations, and therefore validity in one normalisation implies validity in the other, it is likely that the results from Table 1 are more likely accurate, and Table 2 instead shows a large number of false rejections from the J-test.

References

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica*, 1029-1054.

Kleibergen, F., & Paap, R. (2006). Generalized reduced rank tests using the singular value decomposition. *Journal of Econometrics*, 133(1), 97-126.

Lane, S., & Windmeijer, F. (2025). Overidentification testing with weak instruments and heteroskedasticity. Working paper

Newey, W. K., & West, K. D. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55(3), 703–708.

Sargan, J. D. (1958). The estimation of economic relationships using instrumental variables. *Econometrica*, 393-415.

Windmeijer, F. (2021). Testing underidentification in linear models, with applications to dynamic panel and asset pricing models. *Journal of Econometrics*, 105104.

Yogo, M. (2004). Estimating the elasticity of intertemporal substitution when instruments are weak. *Review of Economics and Statistics*, 86(3), 797-810.