Inference with Many Instruments

Michal Kolesár

June 03, 2025

Contents

Summary

The package `ManyIV` implements estimators and confidence intervals in a linear instrumental variables model considered in Kolesár [2018] and Kolesár et al. [2015]. In this vignette, we demonstrate the implementation of these estimators and confidence intervals using a subset of the dataset used in Angrist and Krueger [1991], which is included in the package as a data frame `ak80`. This data frame corresponds to a sample of males born in the US in 1930–39 from 5% sample of the 1980 Census. See `help("ManyIV::ak80")` for details.

Estimation and Inference

The package implements the following estimators via the command `IVreg`

1. Two-stage least-squares (TSLS) estimator
2. Limited information maximum likelihood (LIML) estimator due to Anderson and Rubin [1949].
3. A modification of the bias-corrected two-stage least squares (MBTSLS) estimator (Kolesár et al. [2015]) that slightly modifies the original Nagar [1959] estimator so that it's consistent under many exogenous regressors as well as many instruments, provided the reduced-form errors are homoskedastic.
4. Efficient minimum distance (EMD) estimator (Kolesár [2018]) that is more efficient than LIML under many instrument asymptotics unless the reduced-form errors are Gaussian.

`IVreg` computes the following types of standard errors:

1. Conventional homoskedastic standard errors, as computed by Stata's `ivregress` and `ivreg2`. These standard errors are not robust to many instruments (option `inference=standard`)
2. Conventional heteroskedastic standard errors, as computed by Stata's `ivregress` and `ivreg2`. These standard errors are not robust to many instruments. (option `inference=standard`)
3. Standard errors that are valid under heterogeneous treatment effects as well as heteroskedasticity (labeled `HTE robust`). These standard errors are not robust to many instruments (option `inference=standard`). They are only computed for TSLS and MBTSLS, since LIML is not robust to heterogeneous treatment effects (see Kolesár [2013]).
4. Standard errors based on the information matrix of the limited information likelihood of Anderson and Rubin [1949] (for LIML only). These are not robust to many instruments or heteroskedasticity (option `inference=lil`)
5. Standard errors based on the Hessian of the random-effects likelihood of Chamberlain and Imbens [2004]. These standard errors are for LIML only (since the random-effects ML estimator coincides to LIML), and are robust to many instruments provided the reduced-form errors are Gaussian and homoskedastic (option `inference=re`).
6. Standard errors based on the Hessian of the invariant likelihood (see Kolesár [2018]). These standard errors are for LIML only (since the invariant ML estimator coincides to LIML), and are robust to many instruments provided the reduced-form errors are Gaussian and homoskedastic. This involves some numerical optimization. (option `inference=il`)
7. Many-instrument robust standard errors based on the minimum distance objective function (see Kolesár [2018]) (option `inference=md`). Since the TSLS estimator is not consistent under many-instrument asymptotics, its standard errors are omitted. Unlike the `re` and `il` standard errors, the standard errors for MBTSLS, LIML and EMD do not require the reduced-form errors to be Gaussian, although the homoskedasticity assumption is still needed. In addition, the command computes standard errors for MBTSLS based on the unrestricted minimum distance objective function (`umd`), which allows for treatment effect heterogeneity (provided the reduced-form errors remain homoskedastic), and for failures of the exclusion restriction as considered in Kolesár et al. [2015].

Several of these options may be specified at once:

```
library("ManyIV")
## Specification as in Table V, columns (1) and (2) in
## Angrist and Krueger
IVreg(lwage ~ education + as.factor(yob) | as.factor(qob) *
    as.factor(yob), data = ak80, inference = c("standard",
    "re", "il", "lil"))
#> Call:
#> IVreg(formula = lwage ~ education + as.factor(yob) | as.factor(qob) *
#>     as.factor(yob), data = ak80, inference = c("standard", "re",
#>     "il", "lil"))
#>
#> First-stage F:  4.907069
#>
#> Estimates and standard errors:
#>          Estimate Conventional Conv. (robust) HTE robust        lil         re
#> ols    0.07108105 0.0003390067   0.0003814625         NA         NA         NA
#> tsls   0.08911546 0.0161098202   0.0162120317 0.01760798         NA         NA
```

```
#> liml   0.09287642 0.0177441446   0.0196323640         NA 0.01615829 0.01986004
#> mbtsls 0.09373337 0.0180984698   0.0204147326 0.02223338         NA         NA
#>                il
#> ols            NA
#> tsls           NA
#> liml   0.01978592
#> mbtsls         NA
```

With large data, the md standard errors may take a while to run, as they require estimation of third and fourth moments of the reduced-form errors. In particular, letting M denote the annihilator matrix associated with the matrix (W, Z) of exogenous regressors and instruments, the formulas for these moments require the computation of $\tilde{m}_3 = \sum_{i,j} M_{i,j}^3$, and $\tilde{m}_4 = \sum_{i,j} M_{i,j}^4$. If option `approx=TRUE` is selected (which is the default), to speed up the calculations, the function `ivreg` uses the approximation $\tilde{m}_3 \approx n - 3(k+l)$ and $\tilde{m}_4 \approx n - 4(k+l)$, where n is the sample size, k is the number of instruments, and ℓ is the number of exogenous regressors. This approximation is accurate up to terms of order $O(((k+l)/n)^2)$, and should have a negligible effect on the estimates unless the ratio $(k+l)/n$ is quite large. With this approximation, the calculations are quite fast even for large sample sizes:

```
r1 <- IVreg(lwage ~ education + as.factor(yob) | as.factor(qob) *
    as.factor(yob), data = ak80, inference = "md", approx = TRUE)
print(r1, digits = 4)
#> Call:
#> IVreg(formula = lwage ~ education + as.factor(yob) | as.factor(qob) *
#>     as.factor(yob), data = ak80, inference = "md", approx = TRUE)
#>
#> First-stage F:  4.907069
#>
#> Estimates and standard errors:
#>        Estimate      md     umd
#> liml    0.09288 0.02024      NA
#> mbtsls  0.09373 0.02031 0.01999
#> emd     0.09288 0.02024      NA
```

We can see that the LIML and EMD estimates are identical up to 4 significant digits.

Specification testing

The package also implements two tests for overidentifying restrictions. The first test is the classic Sargan [1958] test. The second test is a modification of the Cragg and Donald [1993] test developed in Kolesár [2018] to make the test robust to many instruments and many exogenous regressors (provided the reduced-form errors are homoskedastic). The command `IVoverid` takes the results of the IV regression as an argument.

```
IVoverid(r1)
#>             statistic   p.value
#> Sargan       25.39429 0.6576361
#> Modified-CD  25.39316 0.6576480
```

Implementation details

Let

$$y_i = x_i\beta + w_i'\delta + \epsilon_i,$$

where $y_i \in \mathbb{R}$ is the outcome variable, $x_i \in \mathbb{R}$ is a single endogenous regressor, $w_i \in \mathbb{R}^\ell$ is a vector of exogenous regressors (covariates), and ϵ_i is a structural error. The parameter of interest is β. In addition, $z_i \in \mathbb{R}^k$ is a vector of instruments.

We observe an i.i.d.~sample $\{y_i, x_i, w_i, z_i\}_{i=1}^n$. Let Y, Z, and W, denote matrices with rows (y_i, x_i), z_i' and w_i'. For any full-rank $n \times m$ matrix A, let $H_A = A(A'A)^{-1}A'$ denote the associated $n \times n$ projection matrix (also known as the hat matrix). Let I_m denote the $m \times m$ identity matrix, and let $Z_\perp = (I_n - H_W)Z$ denote the residual from the sample projection of Z onto W.

Define matrices S and T as in Kolesár [2018]:

$$T = Y'H_{Z_\perp}Y/n, \qquad\qquad S = Y'(I_n - H_{Z,W})Y/(n-k-\ell).$$

Also define $m_{\min}$ and $m_{\max}$ to be the minimum and maximum eigenvalues of the matrix $S^{-1}T$. The estimators TSLS, OLS, MBTSLS, and LIML are all k-class estimators. A k-class estimator estimator with parameter κ is then given by

$$\hat{\beta}(\kappa) = \frac{T_{12} - m(\kappa)S_{12}}{T_{22} - m(\kappa)S_{22}},$$

where $m(\kappa) = (\kappa - 1)(1 - k/n - \ell/n)$. For the estimators above,

$$m_{OLS} = -(1 - k/n - \ell/n) \qquad m_{TSLS} = 0, \qquad m_{MBTSLS} = k/n, \qquad m_{LIML} = m_{\min}.$$

The EMD estimator is not a k-class estimator.

The `li`, `lil`, `re`, and `md` standard errors are based on the formulas described in Kolesár [2018]. In the remainder of this vignette, we briefly describe the formulas for conventional standard errors.

Other standard errors

Stata 13's `ivregress` and `ivreg2` use standard errors for k-class estimators given by

$$\widehat{var}_{\text{Stata}}(\hat{\beta}(\kappa)) = \frac{1}{n}\frac{\hat{\sigma}(\kappa)^2}{T_{22} - m(\kappa)S_{22}},$$

where $\hat{\sigma}(\kappa)^2 = \hat{\epsilon}(\kappa)'\hat{\epsilon}(\kappa)/n$, with $\hat{\epsilon}(\kappa) = y - x\hat{\beta}(\kappa) - W'\hat{\delta}(\kappa)$, and $\hat{\delta}(\kappa) = (W'W)^{-1}W'(y - x\hat{\beta}(\kappa))$. This includes LIML, for which κ is random (Stata disregards that). For OLS, we use the Stata 13 variance estimator $\hat{\sigma} = \hat{\epsilon}_{OLS}'\hat{\epsilon}_{OLS}/(n - \ell - 1)$.

To define the robust standard error estimators, let $\hat{R}_i = Z_{\perp,i}(Z_\perp' Z_\perp)^{-1}Z_\perp x$. Then, for a k-class estimator (including LIML),

$$\widehat{var}_{\text{Stata, robust}}(\hat{\beta}(\kappa)) = \frac{\sum_{i=1}^n \hat{\epsilon}_i(\kappa)^2\hat{R}_i^2}{n^2(T_{22} - m(\kappa)S_{22})^2}.$$

Note that $\widehat{var}_{\text{Stata, robust}}(\hat{\beta}(\kappa))$ and $\widehat{var}_{\text{Stata}}(\hat{\beta}(\kappa))$ don't necessarily converge to the same quantity even under homoskedasticity. For OLS, we use $(n/(n-\ell-1))^{1/2}x_\perp$ in place of $\hat{R}_i$.

One could alternatively use T_{22} in the denominator, or estimate $var(\epsilon_i)$ using $\hat{\sigma}(\beta) = (1, -\beta)S(1, -\beta)'$. Such variance estimators were used in Kolesár et al. [2015]. The alternative denominator makes a big difference, but how we estimate σ^2 matters less.

Other outputs

The first-stage F-statistic reported by `IVreg` is given by

$$F = \frac{n}{k}\frac{T_{22}}{S_{22}}.$$

The Sargan test statistic is given by $nm_{\min}/(1 - p/n - \ell/n + m_{\min})$, and its p-value is based on a χ^2_{k-1} approximation. The Sargan test statistic is based on LIML, unlike in Stata 13's `estat overid`, where it depends on what estimator was used to compute β. The adjusted Cragg-Donald test is described in Kolesár [2018, Section 6].

References

Theodore W. Anderson and Herman Rubin. Estimation of the parameters of a single equation in a complete system of stochastic equations. *The Annals of Mathematical Statistics*, 20(1):46–63, March 1949. doi: 10.1214/aoms/1177730090.

Joshua D. Angrist and Alan B. Krueger. Does compulsory school attendance affect schooling and earnings? *The Quarterly Journal of Economics*, 106(4):979–1014, November 1991. doi: 10.2307/2937954.

Gary Chamberlain and Guido Wilhelmus Imbens. Random effects estimators with many instrumental variables. *Econometrica*, 72(1):295–306, January 2004. doi: 10.1111/j.1468-0262.2004.00485.x.

John G. Cragg and Stephen G. Donald. Testing identifiability and specification in instrumental variable models. *Econometric Theory*, 9(2):222–240, February 1993. doi: 10.1017/S0266466600007519.

Michal Kolesár. Estimation in an instrumental variables model with treatment effect heterogeneity. Working paper, Princeton University, November 2013. URL https://www.princeton.edu/~mkolesar/papers/late_estimation.pdf.

Michal Kolesár. Minimum distance approach to inference with many instruments. *Journal of Econometrics*, 204(1):86–100, May 2018. doi: 10.1016/j.jeconom.2018.01.004.

Michal Kolesár, Raj Chetty, John N. Friedman, Edward Glaeser, and Guido W. Imbens. Identification and inference with many invalid instruments. *Journal of Business and Economic Statistics*, 33(4): 474–484, October 2015. doi: 10.1080/07350015.2014.978175.

Anirudh Lal Nagar. The bias and moment matrix of the general k-class estimators of the parameters in simultaneous equations. *Econometrica*, 27(4):575–595, October 1959. doi: 10.2307/1909352.

John Denis Sargan. The estimation of economic relationships using instrumental variables. *Econometrica*, 26(3):393–415, July 1958. doi: 10.2307/1907619.