Inference with many instruments

Michal Kolesár

2017-07-04

Contents

Summary

The package `ManyIV` implements estimators and confidence intervals in a linear instrumental variables model considered in Kolesár (2017) 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.
4. Efficient minimum distance (EMD) estimator (Kolesár 2017) 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 are computed:

1. Conventional homoscedastic standard errors, as computed by Stata's `ivregress` and `ivreg2`. These standard errors are not robust to many instruments (option `inference=standard`)
2. Conventional heteroscedastic 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 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 (option `inference=lil`)
4. 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 (option `inference=re`).

5. Standard errors based on the Hessian of the invariant likelihood (see Kolesár 2017). 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. This involves some numerical optimization. (option `inference=il`)
6. Many-instrument robust standard errors based on the minimum distance objective function (see Kolesár 2017) (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. 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 homoscedastic), 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:  4.90707
#>
#> Estimates and standard errors:
#>         Estimate Conventional Conv. (robust)       lil      re        il
#> ols    0.0710810  0.000339007    0.000381463        NA      NA        NA
#> tsls   0.0891155  0.016109820    0.016212032        NA      NA        NA
#> liml   0.0928764  0.017744145    0.019632364 0.0161583 0.01986 0.0197859
#> mbtsls 0.0937334  0.018098470    0.020414733        NA      NA        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 annihilaror 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:  4.90707
#>
```

```
#> 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 (2017) to make the test robust to many instruments and many exogenous regressors. The command `IVoverid` takes the results of the IV regression as an argument.

```
IVoverid(r1)
#>             statistic  p.value
#> Sargan        25.3943 0.657636
#> Modified-CD   25.3932 0.657648
```

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 (2017):

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

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 (2017). In the remainder of this vignette, we briefly describe the formulas for conventional standard errors.

Conventional 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),

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

Note that $\widehat{var}_{\text{StataR}}(\hat{\beta}(\kappa))$ 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 (2017 Section 6).

References

Anderson, Theodore W., and Herman Rubin. 1949. "Estimation of the Parameters of a Single Equation in a Complete System of Stochastic Equations." *The Annals of Mathematical Statistics* 20 (1): 46–63.

Angrist, Joshua D., and Alan B. Krueger. 1991. "Does Compulsory School Attendance Affect Schooling and Earnings?" *The Quarterly Journal of Economics* 106 (4): 979–1014.

Chamberlain, Gary, and Guido Wilhelmus Imbens. 2004. "Random Effects Estimators with Many Instrumental Variables." *Econometrica* 72 (1): 295–306.

Cragg, John G., and Stephen G. Donald. 1993. "Testing Identifiability and Specification in Instrumental

Variable Models." *Econometric Theory* 9 (2): 222–40.

Kolesár, Michal. 2017. "Minimum Distance Approach to Inference with Many Instruments."

Kolesár, Michal, Raj Chetty, John Friedman, Edward L. Glaeser, and Guido Wilhelmus Imbens. 2015. "Identification and Inference with Many Invalid Instruments." *Journal of Business & Economic Statistics* 33 (4): 474–84.

Nagar, A. L. 1959. "The Bias and Moment Matrix of the General K-Class Estimators of the Parameters in Simultaneous Equations." *Econometrica* 27 (4): 575–95.

Sargan, J. D. 1958. "The Estimation of Economic Relationships Using Instrumental Variables." *Econometrica* 26 (3): 393–415.